REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

0565

8

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 28, 2000

3. REPORT TYPE AND DATES COVERED
Final - October 1, 1997 - September 30, 2000

4. TITLE AND SUBTITLE
Prediction of Health and Environment Hazards of Chemicals: A Hierarchial Approach Using QMSA and QSAR

5. FUNDING NUMBERS
F49620-98-1-0015

6. AUTHOR(S)
Dr. Subhash C. Basak
Natural Resources Research Institute

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Minnesota
5013 Miller Trunk Highway
Duluth, MN 55811

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
801 North Randolph Street, Room 732
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

During the first year of the project, the majority of effort was spent in the development of novel hierarchial QSAR methods, QMSA techniques and the applications of these methods in the prediction of toxicological, physiochemical and biomedicinal properties of different sets of chemicals. During the second year of the project, our effort was directed towards the development of novel optimal molecular descriptors, the development and use ofnew topological indices, the study of the intercorrelation of a large number of molecular descriptors, and the use of calculated molecular descriptors in the prediction of toxicological and toxicologically-relevant properties. The third year of the project focused on the further expansion of our theoretical molecular descriptor set through the further development of new topological indcies and the acquisition of several other well-known software packages for the calculations of molecualr descriptors, viz., CODESSA v2.0 and MolconnZ-v3.50.

14. SUBJECT TERMS
Chemicals, QMSA, QSAR

15. NUMBER OF PAGES
177

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclas	Unclas	Unclas	

Final Report
of the Air Force Project

Covering research period 8/1/97 to 9/30/00

Agency No: DOD/F49620-98-1-0015
U of M No: 1613-189-6158
NRRI Technical Report No: NRRI/TR-2000/44

Prediction of Health and Environmental Hazards of Chemicals: A Hierarchical Approach Using QMSA and QSAR

September 28, 2000

Submitted by:

Subhash C. Basak, Ph.D.
Principal Investigator
Natural Resources Research Institute
University of Minnesota, Duluth
5013 Miller Trunk Highway
Duluth, MN 55811
Tel: (218)720-4230
Fax: (218)720-4328
Email: sbasak@nrri.umn.edu

The University of Minnesota is an equal opportunity educator and employer.

Table of Contents

Objectives

During the past few years we have been involved in the development of new computational methods for quantifying similarity/dissimilarity of chemicals and applications of quantitative molecular similarity analysis (QMSA) techniques in analog selection and property estimation for use in the hazard assessment of chemicals. We have also explored the mathematical nature of the molecular similarity space in order to better understand the basis of analog selection by QMSA methods. The parameter spaces used for QMSA and analog selection were constructed from nonempirical parameters derived from computational chemical graph theory. Occasionally, graph invariants were supplemented with geometrical parameters and quantum chemical indices to study the relative effectiveness of graph invariants vis-à-vis geometrical and quantum chemical parameters in analog selection and property estimation. We carried out comparative studies of nonempirical descriptor spaces and physicochemical property spaces in selecting analogs. Molecular similarity methods were applied in predicting modes of toxic action (MOA) of chemicals. Our similarity/dissimilarity methods have also found successful applications in the discovery of new drug leads by US drug companies.

In this project, we will have four primary goals: 1) development of a hierarchical approach to molecular similarity, 2) formulation of quantitative structure-activity relationship (QSAR) models for predictive toxicology using a hierarchical approach, 3) applications of hierarchical QSAR and QMSA approaches in computational toxicology related to human health and ecological hazard assessment, and 4) the application of hierarchical QMSA and QSAR approaches in estimating potential toxicity of deicing agents.

The first goal of the project is the use of parameters of gradually increasing complexity, viz., topological, topochemical, geometrical, and quantum chemical indices, in the quantification of molecular similarity/dissimilarity of chemicals. We will take a two-tier approach in this area. First, similarity methods will be used in ordering sets of molecules and in selecting structural analogs of toxic chemicals which pose human health and ecological hazards. Secondly, we will use the properties of selected analogs in estimating toxicologically important properties for chemicals. Although different classes of parameters have been used in the characterization of molecular similarity, no systematic study has been carried out in the use of all four classes of parameters, mentioned above, in analog selection and property estimation. We will apply a hierarchical approach to the use of these four types of theoretical molecular descriptors in the quantification of molecular similarity/dissimilarity.

The second goal consists of the development of hierarchical QSAR models for predicting the toxic potential of chemicals using topological and quantum chemical indices. Initially, we will use parameters calculated by semi-empirical methods such as MOPAC and AMPAC. Parameters calculated by *ab initio* quantum chemical methods will be used in limited cases of QSAR model development, if they are considered necessary.

The third goal of the project will be the prediction of human health hazard and ecotoxicological effects of chemicals using QSAR and QMSA methods developed in the project. Attempts will be made to estimate endpoints, such as, carcinogenicity, mutagenicity, xenoestrogenicity, acute toxicity, transport of chemicals through the blood-brain barrier, biodegradation, and bioconcentration factor.

The fourth goal will involve the utilization of QMSA and QSAR methods developed as part of this project in predicting the potential toxicity of deicing agents.

Status of Efforts

During the first year of the project the majority of effort was spent in the development of novel hierarchical QSAR methods, QMSA techniques and the applications of these methods in the prediction of toxicological, physicochemical and biomedicinal properties of different sets of chemicals. Our dissimilarity methods were used to group JP-8 constituents into a small number of clusters that can be used in selecting surrogate mixtures for JP-8 in the Air Force's toxicological studies. The clustering was done using algorithmically derived molecular descriptors calculated by our computer program POLLY. Such parameters can be calculated for any molecular structure, real or hypothetical. This makes the clustering methods independent of any experimentally determined property of the JP-8 constituents.

During the second year of the project, our effort was directed towards the development of novel optimal molecular descriptors, the development and use of new topological indices, the study of the intercorrelation of a large number of molecular descriptors, and the use of calculated molecular descriptors in the prediction of toxicological and toxicologically-relevant properties. We also explored the possibility of developing integrated QSAR (I-QSAR) with the combination of chemodescriptors derived from computational chemistry and biodescriptors derived from biological techniques such as proteomics.

The third year of the project has focused on the further expansion of our theoretical molecular descriptor set through the further development of new topological indices and the acquisition of several other well-known software packages for the calculation of molecular descriptors, *viz.*, CODESSA v2.0 and Molconn-Z v3.50. Along with this expansion, we have continued our pioneering studies in the intercorrelation of large molecular descriptor sets and the use of this expanded descriptor set in the prediction of toxicological and toxicologically-relevant properties. We have also begun the initial exploration of the creation of biodescriptors, derived from matrix invariants, to handle data from proteomics maps and have developed several new methods for the characterization of DNA sequences.

Accomplishments/ New Findings

The following is the summary of accomplishments of the various tasks of the project during the reporting period:

Task 1: Development of Databases

Years 1 & 2 Databases of toxicological endpoints and physicochemical properties have been developed from published literature. Such data have been used in the hierarchical QSAR and QMSA studies (vide infra).

Year 3 Efforts to develop more databases from published literature have tapered off, with more emphasis being placed on other aspects of the project. However, we have been making efforts to acquire a number of large, proprietary databases from various companies for the purposes of testing some of our methods against "real" drug-development databases.

Task 2: Development of a Comprehensive Computer Program for Calculating Topological Molecular Descriptors

Years 1 & 2 POLLY can calculate more than one hundred topological indices (TIs). We have been working to develop algorithms to calculate other topological descriptors such as local invariants. Such indices will be tested in hierarchical QSAR and QMSA research.

Year 3 A new software module associated with POLLY has been developed and is currently being tested. This module, called TRIPLET, can calculate 100 local vertex invariants (LOVIs) which are also known as triplet indices.

Task 3: Integration of Graph Theory and Quantum Chemistry for QSAR

Years 1 & 2 Ongoing research in this area focused on the use of weighted graphs, pseudographs in the development of novel descriptors. This will lead to novel invariants that can encode information not quantified by existing molecular descriptors. In the second year of the project, a paper was submitted for publication that studied the interrelationship of over 200 topological indices.

Year 3 The intercorrelation study submitted last year was published this spring in the *Journal of Chemical Information and Computer Science* (Basak et al. 2000). This study is being followed with a more rigorous study involving using a larger set of 318 indices on an expanded set of databases. Additionally, our findings that in many cases quantum chemical indices do no better than topological indices in QSAR modeling are being borne out by the work of other researchers.

Task 6: Characterization of Structure Using Theoretical Structural Descriptors

Years 1 & 2 We have used topological indices and principal components (PCs) derived from them in the characterization of a set of isospectral graphs which cannot be discriminated by the eigenvalues of the adjacence matrix of molecular graphs. This result was published in the *Journal of Chemical Information and Computer Sciences* (Balasubramanian and Basak 1998).

Attempts have been made to devise descriptors that characterize chemical structures optimally. This has been done through the use of weighted graphs. Invariants based on line graphs have also been used for QSAR studies. Both of these techniques involve the development of novel descriptors for the characterization of molecular structure.

Year 3 Work on optimized molecular descriptors with Dr. Randic has continued, resulting in a number of new publications. Additionally, this work has spread into new fields with our development of methods to characterize protein structure and folding through the use of novel invariants.

Task 7: Development of Hierarchical QMSA Models

Years 1 & 2 Topostructural, topochemical, geometrical as well as quantum chemical parameters have been used in the development of QMSA methods. We carried out a dissimilarity-based clustering of JP-8 constituents into fourteen clusters. A mixture of compounds selected from each cluster can be used as surrogates for the complex JP-8 mixture.

The method has also been used in the clustering of a large, virtual, combinatorial library of Psoralen derivatives. The results of this analysis were presented in five papers at the International Biophysics Congress, New Delhi, September 19-23, 1999.

Year 3 Additional studies involving the development and refinement of the hierarchical QMSA method were presented at the Second Indo-US Workshop on Mathematical Chemistry, Duluth, MN, May 30-June 3, 2000 and at the National American Chemical Society meeting, Washington, D.C., August 20-24, 2000.

Task 8: Development of Hierarchical Approach to QSAR

Year 1 & 2 Quantum chemical parameters calculated by semiempirical methods have been used in hierarchical QSAR models for predicting toxicity and toxicologically relevant physicochemical properties. Several manuscripts have been published in peer-reviewed journals.

Our hierarchical approach has been used in the development of QSAR models for the prediction of toxicity (e.g., aquatic toxicity, LC_{50}, of a set of benzene derivatives, skin penetration by polycyclic aromatic hydrocarbons, mutagenicity, etc). We have used mainly linear statistical methods such as variable clustering, principal components analysis, etc, for model building. In the area of neural net analysis, we used linear as well as nonlinear methodology. In the case of toxicity of benzene derivatives, there were some improvements in the model over the linear statistical methods by the applications of neural net methodology.

Year 3 Findings of recent hierarchical QSAR modeling studies were presented at both the Second Indo-US Workshop on Mathematical Chemistry and at the National American Chemical Society meeting. We have continued working to examine the relative effectiveness of linear and non-linear statistical methods versus linear and non-linear neural network methods and has resulted in the publication of two manuscripts and the submission of two other studies for peer-review and publication.

Work on the development of novel biodescriptors has been progressing well. Our collaborative efforts aim at the development of a series of novel invariants for the characterization of proteomics maps. We hope to continue these studies to move beyond the theoretical stage to develop software to calculate these invariants and to test them in QSAR model development.

Personnel Supported

Subhash C. Basak, Principal Investigator
Alexandru Balaban, Visiting Scientist (Distinguished Professor; Polytechnic Univ., Bucharest, Romania)
Krishnan Balasubramanian, Consultant (Professor; Arizona State Univ., Tempe, AZ)
Doug Dilla, Undergraduate Student
Jassen Dagit, Undergraduate Student
Greg Grunwald, Applications Programmer
Brian Gute, Assistant Scientist
Douglas Hawkins, Co-principal Investigator (Chairman – Dept. of Applied Statistics; Univ. of Minnesota, St. Paul, MN)
Keith B. Lodge, Co-principal Investigator
Denise Mills, Technician
Ashesh Nandy, Visiting Scientist (Indian Institute of Chemical Biology, Calcutta, India)
Sonja Nikolic, Distinguished Professor (Rugjer Boskovic Institute, Zagreb, The Republic of Croatia)
David W. Opitz, Consultant (Assistant Professor; Univ. of Montana, Missoula, MT)
Milan Randic, Visiting Scientist (Distinguished Professor; Drake Univ., Des Moines, IA)
Xiaofang Shi, Graduate Student (Dept. of Applied Statistics, Univ. of Minnesota, St. Paul, MN)
Nenad Trinajstic, Visiting Scientist (Distinguished Professor; Rugjer Boskovic Institute, Zagreb, The Republic of Croatia)
Marjan Vracko, Visiting Scientist (Professor; National Institute of Chemistry, Slovenia)

Publications

The following peer-reviewed papers, which are currently either published, in press, or submitted, report results of research carried out between August 1, 1997 and September 30, 2000.

1997 Characterization of molecular structures using topological indices, S.C. Basak and B.D. Gute, *SAR QSAR Environ. Res.*, **7**, 1-21 1997.

Computational study of the environmental fate of selected aircraft fuel system deicing compounds, G.W. Mushrush, S.C. Basak, J.E. Slone, E.J. Beal, S. Basu, W.M. Stalick and D.R. Hardy, *J. Environ. Sci. Health*, A32, 2201-2211, 1997.

Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, B. D. Gute and S. C. Basak, *SAR QSAR Environ. Res.*, **7**, 117-131, 1997.

1998 Characterization of isospectral graphs using graph invariants and derived orthogonal parameters, K. Balasubramanian and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **38**, 367, 1998.

Characterization of the molecular similarity of chemicals using topological invariants, S. C. Basak, B. D. Gute, and G. D. Grunwald, in: *Advances in Molecular Similarity*, JAI Press, pp. 171-185,vol. 2, R. Carbo-Dorca and P. G. Mezey (Eds), 1998.

The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, S. C. Basak, B. D. Gute and G. D. Grunwald, In *QSAR in Environmental Sciences - VII*, F. Chen and G. Schuurmann, eds., SETAC Press, Pensacola, FL, 1998, Chapter 17, p 245-261.

1999 A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, S.C. Basak, B.D. Gute and G.D. Grunwald, In *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A.T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, p 675-696.

Assessment of the mutagenicity of chemicals from theoretical structural parameters: A hierarchical approach, S.C. Basak, B.D. Gute, and G.D. Grunwald, *SAR QSAR Environ. Res.*, **10**, 117-129, 1999.

Correlation between structure and normal boiling point of acyclic carbonyl compounds, A. T. Balaban, D. Mills and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **39**, 758-764, 1999.

Hazard assessment modeling: An evolutionary ensemble approach, D.W. Opitz, S.C. Basak and B.D. Gute, In: *Genetic and Evolutionary Computation*, Eds. W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, & R.E. Smith, Morgan Kaufmann: San Francisco, 1999, p 1643-1651.

Information theoretic indices of neighborhood complexity and their applications, S.C. Basak, In *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A.T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, p 563-593.

Normal boiling points of 1,ω-alkanedinitriles: The highest increment in a homologous series, A.T. Balaban, S.C. Basak and D. Mills, *J. Chem. Inf. Comput. Sci.*, **39**, 769-774, 1999.

Optimal molecular descriptors based on weighted path numbers, M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **39**, 261-266, 1999.

Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters, S.C. Basak, B.D. Gute, and S. Ghatak, *J. Chem. Inf. Comput. Sci.*, **39**, 255-260, 1999.

Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): a hierarchical QSAR approach, B. D. Gute, G. D. Grunwald, and S. C. Basak, *SAR. QSAR Environ. Res.*, **10**, 1-15, 1999.

Use of statistical and neural net methods in predicting toxicity of chemicals: A hierarchical QSAR approach, S.C. Basak, B.D. Gute, G.D. Grunwald, D.W. Opitz and K. Balasubramanian, In *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools* - Papers from the 1999 AAAI Symposium, March 22-24, 1999, Stanford, CA, TR SS-99-01, AAAI Press: Menlo Park, CA, 1999, p 108-111.

2000 A comparative QSAR study of benzamidines complement-inhibitory activity and benzene derivatives acute toxicity, S.C. Basak, B.D. Gute, B. Lucic, S. Nikolic and N. Trinajstic, *Computers & Chemistry*, **24**, 181-191, 2000.

Construction of high-quality structure-property-activity regressions: The boiling points of sulfides, M. Randic and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **40**, 899-905, 2000.

Multiple regression analysis with optimal molecular descriptors, M. Randic and S.C. Basak, SAR QSAR Environ. Res., **11**, 1-23, 2000.

On 3-D graphical representation of DNA primary sequences and their numerical characterization, M. Randic, M. Vracko, A. Nandy and S. C. Basak, *J. Comput. Chem.*, **40**, 1235-1244, 2000.

QSPR modeling: Graph connectivity indices versus line graph connectivity indices, S. C. Basak, S. Nikolic, N. Trinajstic, D. Amic and D. Beslo, *J. Chem. Inf. Comput. Sci.*, **40**, 927-933, 2000.

Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, A. Nandy and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **40**, 915-919, 2000.

Topological indices: Their nature and mutual relatedness, S. C. Basak, A. T. Balaban, G. D. Grunwald and B. D. Gute, *J. Chem. Inf. Comput. Sci.*, **40**, 891-898, 2000.

Use of graph invariants in QMSA and predictive toxicology, S.C. Basak and B.D. Gute, In *Discrete Mathematical Chemistry*, Eds. P. Hansen, P. Fowler, M. Zheng, DIMACS Series 51, American Mathematical Society: Providence, Rhode Island, 2000, pages 9-24.

Use of statistical and neural net approaches in predicting toxicity of chemicals, S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian and D. Opitz, *J. Chem. Inf. Comput. Sci.*, **40**, 885-890, 2000.

In press

Molecular similarity based estimation of properties: A comparison of structure spaces and property spaces, B.D. Gute, G.D. Grunwald, D. Mills and S.C. Basak, *SAR QSAR Environ. Res.*, 2000.

On characterization of physical properties of amino acids, M. Randic, D. Mills and S. C. Basak, *Int. J. Quant. Chem.*, 2000.

On ordering of folded structures, M. Randic, M. Vracko, M. Novic and S. C. Basak, *Mathematical Chemistry, MATCH*, 2000.

Quantitative comparison of five molecular structure spaces in selecting analogs of chemicals, S.C. Basak, B.D. Gute, and G.D. Grunwald, *Mathl. Model. Comput. Sci.*, 2000.

Reverse Wiener index, A. T. Balaban, D. Mills and S. C. Basak, *Croat. Chim. Acta*, 2000.

Use of mathematical structural invariants in analysing combinatorial libraries: A case study with Psoralen derivatives, S.C. Basak, D. Mills, B.D. Gute, A.T. Balaban, K. Basak and G.D. Grunwald, In *Some Aspects of Mathematical Chemistry*, Eds. D.K. Sinha, S.C. Basak, R.K. Mohanty and I.N. Basumallick, Visva-Bharati University: Santiniketan, West Bengal, India, 2000.

Variable molecular descriptors, M. Randic and S.C. Basak, In *Some Aspects of Mathematical Chemistry*, Eds. D.K. Sinha, S.C. Basak, R.K. Mohanty and I.N. Basumallick, Visva-Bharati University: Santiniketan, West Bengal, India, 2000.

Accepted

Modelling the solubility of aliphatic alcohols in water. Graph connectivity indices versus line graph connectivity indices, S. Nikolic, N. Trinajstic, D. Amic, D. Beslo and S. C. Basak, In *QSAR/QSPR Studies by Molecular Descriptors*, M. V. Diudea, Ed., Nova Science Publishers, New York, USA, 2000.

Submitted

A neural net-based QSAR algorithm (PCANN) and its comparison with hologram- and multiple linear regression-based QSAR approaches applied to 1,4-dihydropyridine-based calcium channel antagonists, V.N. Viswanadhan, G.A. Mueller, S.C. Basak and J.N. Weinstein, *J. Chem. Inf. Comput. Sci.*, 2000.

A new descriptor for structure-property and structure-activity correlations, M. Randic and S.C. Basak, *J. Chem. Inf. Comput. Sci.*, 2000.

A novel 2-D graphical representation of DNA sequences of low degeneracy, X. Guo, M. Randic and S.C. Basak, *Chem. Phys. Lett.*, 2000.

Characterization of DNA primary sequences based on the average distances between bases, M. Randic and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, 2000.

Distance indices and their hyper-counterparts: Intercorrelation and use in the structure-property modeling, N. Trinajstic, S. Nikolic, S.C. Basak and I. Lukovits, *SAR QSAR Environ. Res.*, 2000.

On structural interpretation of distance related topological indices, M. Randic, A.T. Balaban and S.C. Basak, *J. Chem. Inf. Comput. Sci.*, 2000.

On the characterization of DNA primary sequences by triplet of nucleic acid bases, M. Randic, X. Guo and S.C. Basak, *J. Chem. Inf. Comput. Sci.*, 2000.

On use of the variable connectivity index $^1\chi^f$ in QSAR:Toxicity of aliphatic ethers, M. Randic and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, 2000.

Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach, S.C. Basak, D.R. Mills and A.T. Balaban, *J. Chem. Inf. Comput. Sci.*, 2000.

QSAR with few compounds and many features, D.M. Hawkins, S. C. Basak and X. Shi, *J. Chem. Inf. Comput. Sci.*, 2000.

Copies of manuscripts published since the 1999 year-end report are attached as Appendix 1. Copies of the manuscripts at various levels of review and publication have been omitted for the sake of brevity.

Interactions/ Transitions

Transitions

1. Applied computational methods in the design of a set of six anti-epileptic carbamates by Professor Alexandru T. Balaban, Vice President, Rumanian Academy of Sciences.

2. Worked with Dr. James Riviere, North Carolina State University, in the clustering of JP-8 components using dissimilarity methods developed at NRRI.

3. Worked with Dr. Alexander Gybin, The Chormaline Corporation, Duluth, MN in the computer-assisted design of photoactive chemicals

4. Applied computational methods in the design of a set of novel photoactive chemicals by Professor Alexandru T. Balaban, Vice President, Rumanian Academy of Sciences (with Dr. Alexander Gybin, Chormaline Corporation, Duluth, MN).

5. Worked with Dr. Frank Witzmann, IUPUI, in the development of integrated QSAR methods using chemodescriptors and biodescriptors.

6. Worked with Dr. Hirak Basu, SLIL Technology, Madison, WI, to generate a virtual library of about 80,000 chemicals to carry out dissimilarity based design of novel anticancer drugs using POLLY parameters.

7. Worked with Dr. Marjan Vracko, National Institute of Chemistry, Ljubljana, Slovenia, to apply our hierarchical QSAR approach to predict the toxicity of chemicals of interest to the European community.

8. Currently working on a long-term collaborative project with Dr. Indira Ghosh, Astra/Zeneca, Bangalore, India, to implement and use topological indices for clustering and analysis of their large, proprietary databases for the discovery of novel lead compounds.

Meetings/ Seminars/ Invited Presentations

1. Dr. S.C, Basak was the Co-Chairperson of the First Indo/US Workshop on Mathematical Chemistry, organized jointly by NRRI and Visva Bharati University, Santiniketan, West Bengal India, Jan 9-13, 1998. Basak presented the following papers at the workshop:
   i. *Graph invariants, molecular similarity and QSAR* co-authored by B.D. Gute and G.D. Grunwald.
   ii. *Weighted paths as novel optimal molecular descriptors* authored jointly by M. Randic and Basak.
   iii. *The utility of hierarchical model development in examining the structural basis of properties* authored by B.D. Gute, G.D. Grunwald and Basak.
   iv. *Weighted K-nearest neighbors property estimation in molecular similarity* authored by G.D. Grunwald, B.D. Gute and Basak.
   v. *Dissimilarity based clustering of psoralen derivatives in the topological structure space: A strategy for drug design* authored by Basak, G.D. Grunwald, D. Panja, K. Basak and B.D. Gute.

2. Dr. S.C. Basak gave several invited lectures at various national and international symposia during his stay in India from December 23, 1997 through January 31, 1998. These lectures included:
   i. A distinguished lecture *Rational drug design and Ayurvedic medicine* at the conference organized by the Association of Ayurvedic Doctors of India (AADI), January 4, 1998.
   ii. An invited lecture on *Use of computational methods and Ayurvedic knowledge in modern drug discovery* at the conference AYURVEDA TODAY, January 8, 1998.
   iii. An invited seminar on *Assessment of genotoxicity of chemicals from structure: A computational approach* at the Annual Conference of the Indian Association for Cancer Congress, Calcutta, January 21-24, 1998. The lecture was co-authored by B.D. Gute and G.D. Grunwald.

3. Dr. S.C. Basak chaired a session at the DIMACS Workshop on Discrete Mathematical Chemistry, March 23-25, 1998, held at Rutgers University, New Jersey. He also presented an invited paper entitled *Use of graph invariants in QSAR and predictive toxicology* at the conference authored jointly by Basak, B.D. Gute and G.D. Grunwald.

4. Dr. S.C. Basak gave an invited presentation entitled *A computational approach to predicting toxicity: Possible applications to JP8 jet fuel* at the First International Conference on the Environmental Health and Safety of Jet Fuels, organized jointly by US Air Force, National Institute of Occupational Safety and Health, USEPA National Exposure Research Laboratory and American Industrial Hygiene Association, April 1-3, 1998, San Antonio, TX.

5. Dr. S.C. Basak presented the following papers at the International Conference Computational Methods in Toxicology held April 20-22, 1998, Dayton, OH:
   i. *Use of computational methods in predicting potential toxicity of chemicals* authored jointly by Basak, B.D. Gute and G.D. Grunwald.
   ii. *On construction of optimal molecular descriptors* authored jointly by M. Randic and Basak.

iii. *Predicting mode of action of chemicals from structure: A hierarchical approach* authored jointly by Basak, G.D. Grunwald and B.D. Gute.

iv. *A hierarchical approach to predictive toxicology using computed molecular descriptors* authored jointly by B.D. Gute, G.D. Grunwald and Basak

6. Dr. S.C. Basak presented a paper *Dissimilarity-based clustering of psoralen derivatives in the topological structure space: A strategy for drug design* at the Second Annual Chemoinformatics Workshop, organized by the Cambridge Health Institute, Boston, MA, June 15-16, 1998. The paper was co-authored by G. D Grunwald and B.D. Gute.

7. Dr. S.C. Basak presented an invited seminar *Novel drug design methods: Assessing activity and toxicity using computational chemistry* at the Department of Molecular Biology and Genetics, University of Guelph, Ontario, Canada, July 3, 1998.

8. Dr. S.C. Basak presented the invited lecture *Use of theoretical structural descriptors in molecular design and hazard assessment of chemicals* to the scientists of the computer-aided drug design company NANODESIGN, INC, Toronto, Canada, July 6, 1998.

9. Dr. S.C. Basak attended the First Environmental Management Science Program Workshop organized jointly by the American Chemical Society and the Office of Environmental Management, Department of Energy, Chicago, IL, July 27-30, 1998.

10. Dr. S.C. Basak presented the invited lecture *Theoretical molecular descriptors for the prediction of bioactivity /toxicity, selection of analogs, discovery and optimization of leads* authored jointly by Basak, B.D. Gute, G.D. Grunwald and A.T. Balaban at the Astra Symposium on Advances in Medicinal Chemistry organized by the Astra company, Bangalore, September 17-19, 1998.

11. Dr. S.C. Basak presented the invited lecture *Prediction of bioactivity of chemicals from structure: A computational approach* at the Indian Institute of Science, Bangalore, India, September 20, 1998.

12. Dr. S.C. Basak presented the invited lecture *Integration of traditional Indian medicine and chemoinformatics for rapid drug discovery* at the conference organized jointly by East India Pharmaceutical Company, Calcutta, October 12, 1998.

13. B.D. Gute presented an invited talk *A hierarchical QSAR approach to predicting carcinogenicity of chemicals* authored jointly, by S.C. Basak, Gute and G.D. Grunwald, at the 19th Annual Society of Environmental Toxicology and Chemistry meeting, Charlotte, North Caroline, November 15-19, 1998.

14. Dr. S.C. Basak presented the invited lecture *Clustering of JP-8 constituents into structurally dissimilar groups: A novel computational strategy for predictive toxicology* authored jointly by Basak and G.D. Grunwald, at the Air Force Office of Scientific Research JP-8 Jet Fuel Toxicology Workshop, held at the University of Arizona, Tucson, AZ, December 2-3, 1998.

15. Dr. S.C. Basak presented the invited lecture on *Novel drug discovery methods: Predicting pharmacological and toxicological properties of chemicals using computational chemistry* at the Meharry Medical College, Nashville, TN, January 19, 1999.

16. Dr. S.C. Basak delivered the first distinguished lecture in Mathematical Chemistry on *From graph invariants to molecular design: 25 years after the connectivity index* at Visva Bharati University, Santiniketan, West Bengal, India, February 11, 1999.

17. Dr. S.C. Basak presented the invited seminar *Theoretical molecular descriptors for the prediction of bioactivity, toxicity, selection of analogs, discovery and optimization of leads* at the Wockhardt Research Centre, Aurangabad, Maharashtra, India, on February 15, 1999.

18. Dr. S.C. Basak presented the invited lecture *Prediction of bioactivity of chemicals from structure: A hierarchical computational approach* at Bharatiya Vidya Bhavans Swami Prakashananda Ayurvedic Research Center, Mumbai, India, on February 18, 1999.

19. Dr. S.C. Basak presented the invited lecture on *Toxicology in silico: Addressing the quagmire of environmental pollution and protecting public health using computational chemistry* authored jointly by Basak, B.D. Gute, David Opitz and G.D. Grunwald at the International Symposia Series: Reducing the Environmental Impacts of Toxic Chemicals in Asian Economies. The Impacts of Toxic Chemicals and Pollutants on Public Health, the Ecology and the Environment of the Bengal Basin - Bangladesh and India, Dhaka Bangladesh, on March 1, 1999.

20. Dr. S.C. Basak presented the invited seminar on *Novel drug discovery methods: Predicting pharmacological and toxicological properties of chemicals using computational chemistry* at the School of Pharmacy, Dhaka University, Dhaka, Bangladesh on March 4, 1999.

21. Dr. S.C. Basak presented the invited talk *Computational toxicology: A cost effective approach for the protection of human and environmental health* at the International Conference at Santiniketan, India, March 7, 1999.

22. Dr. S.C. Basak gave the invited presentation *Estimation of DNA damage from toxic chemicals by graphical techniques* authored jointly by A. Nandy, C. Raychaudhury, S. Ghosh, and Basak on March 8, 1999.

23. Dr. S.C. Basak attended the at the International Conference Smarter Lead Optimization: Easing the Bottleneck organized by Cambridge Health Institute, March 18-19, 1999, San Diego, CA and gave the following presentations:
    i. *A computational approach to predicting toxicity and toxic modes of action of chemicals from structure.*
    ii. *Topological indices as molecular descriptors for lead optimization* authored jointly by A.T. Balaban and Basak.

24. Dr. S.C. Basak attended the American Association of Artificial Intelligence conference, Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, Stanford University, March 22-24, 1999 to present the following lectures:
    i. *Use of statistical and neural net methods in predicting toxicity of chemicals: A hierarchical QSAR approach* authored jointly by Basak, G.D. Grunwald, B.D. Gute, K. Balasubramanian and D. Opitz.
    ii. *A Graphical Technique for Preliminary Assessment of Effects on DNA Sequences from Toxic Substances* authored jointly by A. Nandy, C. Raychaudhury and Basak.

25. Dr. Basak presented the following papers at the QSAR Gordon Conference, July 25-30, 1999, Tilton, New Hampshire:
    i. *A hierarchical QSAR approach for predicting property/activity of chemicals* authored by Basak, G.D. Grunwald, B.D. Gute, D. Mills, K. Balasubramanian and A.T. Balaban.
    ii. *Topological indices as molecular descriptors for QSAR* authored by A.T. Balaban and Basak.

26. On a trip to Europe and India during September of 1999, Dr. S.C. Basak gave the following invited presentations:
    i. *A hierarchical qsar approach for predicting property/activity of chemical from structure* at the Rugjer Boskovic Institute, Zagreg, The Republic of Croatia.
    ii. *Predicting property/activity/toxicity of chemicals from structure: A hierarchical QSAR approach* at the National Institute of Chemistry, Slovenia.
    iii. *Prediction of activity/toxicity of chemicals from structure using graph invariants* at the Visva Bharati University, Santiniketan, West Bengal, India.
    iv. *Predicting biomedicinal and toxicological properties of chemicals using molecular descriptors* at the University of Delhi, India.
    v. *The utility of Ayurvedic medicine for modern drug discovery: An exploratory analysis* at the conference organized by the East India Pharmaceutical Company, Calcutta.

27. During his trip to India in September of 1999, Subhash Basak also attended the 13th International Biophysics Congress, New Delhi, and presented the following papers:

i. *Clustering of Psoralen derivatives using topological invariants: A strategy for molecular design* coauthored by G.D. Grunwald, A.T. Balaban and K. Basak.

ii. *A hierarchical QSAR approach to predicting bioactivity of chemicals using theoretical molecular descriptors* coauthored by B.D. Gute, D. Mills, G.D. Grunwald, D. Opitz and K. Balasubramanian.

iii. *Modeling the solubility of aliphatic alcohols in water, graph connectivity indices versus line graph connectivity indices* coauthored by D. Amic, S. Nikolic, N. Trinajstic and D. Beslo.

iv. *Design of high quality structure-property regressions* coauthored by M. Randic.

v. *On numerical characterization of DNA primary sequences* coauthored by M. Randic, M. Vracko and A. Nandy.

28. Dr. Basak gave an invited presentation on *Development of hierarchical qsar models for predicting toxicity of chemicals: Statistical and neural net approaches* at the Air Force Predictive Toxicology Conference, Wright Patterson Air Force Base, Dayton, OH.

29. Subhash Basak gave an invited presentation *Exploring the scientific basis of Ayurvedic medicine: A computational approach* at the conference *Beyond Conventional Healthcare: Understanding Alternative Choices* organized by the University of Wisconsin, Superior, Nov., 1999.

30. Dr. Basak participated in the 1999 *Partners in Environmental Technology Symposium and Workshop* held in Washington, D.C.

31. Subhash Basak presented the invited lecture *Applications of theoretical molecular descriptors in drug discovery and predictive toxicology: A computational approach* at the University of Montana, Missoula.

32. Dr. Basak gave the invited presentation *Clustering of JP-8 chemicals using structure spaces and property spaces: A computational approach* authored jointly by B.D. Gute, G.D. Grunwald, D. Mills, J. Riviere and D. Opitz at the Air Force Office of Scientific Research *JP-8 Jet Fuel Toxicology Workshop*, University of Arizona, Tucson, Jan., 2000.

33. Subhash Basak gave the following invited lectures/ presentations during his trip to India, Feb., 2000:

    i. *Predicting biomedical and toxicological properties of chemicals using molecular descriptors: A hierarchical QSAR approach* at the *International Conference on Medicinal Chemistry and Biocatalysis* organized by Delhi University. He also presented the following four posters in the same conference:

       (a) *Clustering of JP-8 chemicals using structure spaces and property spaces: A computational approach* authored jointly by Basak, B.D. Gute, G.D. Grunwald, D. Mills, J. Riviere and D. Opitz.

       (b) *Prediction of gas chromatographic retention indices using variable connectivity index* authored jointly by M. Randic, Basak, M. Pompe and M. Novic.

       (c) *Clustering of Psoralen derivatives using topological invariants: A strategy for molecular design* authored jointly by Basak, D. Mills, A.T. Balaban, K. Basak and G.D. Grunwald.

       (d) *A novel structure-activity approach to benzamidines complement inhibitory activity* authored jointly by Basak, B. Lucic, S. Nikolic and N. Trinajstic.

    ii. Basak also gave the invited presentation *Applications of theoretical molecular descriptors in drug discovery and predictive toxicology: A computational approach* at the Ranbaxy Research Laboratories, Udyog Vihar Industrial Area, Gurgaon, Hariyana, India.

34. D. Mills presented the paper *On the use of variable connectivity index for characterization of amino acids*, co-authored by Basak and M. Randic, at the *40th Sanibel Symposium on Atomic, Molecular, Biophysical and Condensed Matter Theory* organized by the Quantum Theory Project, at the University of Florida, March 2000.

35. Dr. Basak gave the presentation *Estimating physicochemical and toxicological properties of chemicals from calculated molecular descriptors* co-authored by D. Mills, B.D. Gute, D. Opitz and K. Balasubramanian at the Dept. of Energy's *Environmental Management Sciences Program National Workshop* in Atlanta, April, 2000.

36. Subhash Basak gave the lecture *Predicting property/activity/toxicity of chemicals using calculated molecular descriptors* at the University of Florida, Gainesville.

37. Dr. Basak and co-workers presented the following papers at the Second Indo-US Workshop on Mathematical Chemistry, organized by NRRI and Visva Bharati University, India:
    i. A.T. Balaban presented the poster *On the clustering of Psoralens* co-authored by Basak, D. Mills, K. Basak, and G.D. Grunwald.
    ii. B.D. Gute presented the poster *Molecular similarity-based estimation of properties: A comparison of structure spaces and property spaces* co-authored by G.D. Grunwald, D. Mills and S.C. Basak.
    iii. Dr. Basak presented the invited lecture *A hierarchical QSAR approach for predicting property/activity/toxicity of chemicals using theoretical structural descriptors* co-authored by B.D. Gute, D. Mills, A.T. Balaban, D. Opitz and K. Balasubramanian.
    iv. B.D. Gute presented the poster *Clustering of chemical using theoretical structure spaces: A case study with 476 diverse chemicals* co-authored by Basak, G.D. Grunwald and D. Mills.
    v. D. Mills presented the poster *Clustering of JP-8 chemicals using property spaces and structure spaces: A novel tool for hazard assessment* co-authored by Basak, G.D. Grunwald, B.D. Gute and J.E. Riviere.
    vi. M. Randic presented the poster *On use of the variable connectivity index $^1\chi^f$ in QSAR: Toxicity of aliphatic ethers* co-authored by Basak.
    vii. A.T. Balaban presented the invited lecture *Topological indices as valuable molecular descriptors for QSAR and QSPR* co-authored by O. Ivanciuc, D. Mills and Basak.
    viii. M. Pompe presented the poster *Prediction of gas chromatographic retention indices for oxygenated compounds using variable connectivity index $^1\chi^f$* co-authored by M. Veber, M. Randic, M. Novic and Basak.
    ix. A.T. Balaban presented the poster *Topological indices: Their nature and mutual relatedness* co-authored by Basak, G.D. Grunwald and B.D. Gute.

38. Dr. Basak and collaborators made the following presentations at the American Chemical Society Annual meeting recently in Washington, D.C.:
    i. A.T. Balaban presented the invited lecture *Trends and possibilities for future developments of topological indices* authored jointly by Balaban and S.C. Basak.
    ii. B.D. Gute presented the invited lecture *Use of graph invariants for the prediction of property/activity/toxicity of chemicals* authored jointly by S.C. Basak, Gute, D. Mills and A.T. Balaban.
    iii. Dr. Basak presented the lecture *Similarity-based estimation of properties: A comparison of structure spaces* authored jointly by B.D. Gute, G.D. Grunwald, D. Mills and S.C. Basak.
    iv. D. Mills presented the poster *Clustering of JP-8 chemicals using structure spaces and property spaces: A computational approach* authored jointly by Mills, S.C. Basak, G.D. Grunwald, B.D. Gute and J. Riviere.
    v. D. Mills presented the poster *Hierarchical clustering of Psoralen derivatives using topological invariants: A strategy for molecular design* authored jointly by Mills, S.C. Basak, B.D. Gute, A.T. Balaban, G.D. Grunwald and K. Basak.
    vi. D. Mills presented the poster *Use of variable connectivity indices on biological molecules* authored jointly by Mills, M. Randic and S.C. Basak.

39. Dr. Basak visited Milan, Italy (early September 2000) to discuss collaborative projects with colleagues at the Istituto di Ricerche Farmacologiche "Mario Negri" and Milan Chemometric Research Group, Department of Environmental Sciences. He traveled to Slovenia and Croatia, to develop and discuss joint quantitative structure-activity/toxicity/property relationship (QSAR/ QSPR/ QSTR) research papers and projects with colleagues at the National Institute of Chemistry, Ljubljana, Slovenia and the Rugjer Boskovic Institute.

Honors and Awards

1. Dr. S.C. Basak was the Co-Chairperson of the *First Indo-US Workshop on Mathematical Chemistry*, organized jointly by NRRI and Visva Bharati University, Santiniketan, West Bengal India, Jan 9-13, 1998.

2. Dr. S.C. Basak chaired a session at the *DIMACS Workshop on Discrete Mathematical Chemistry*, March 23-25, 1998, held at Rutgers University, New Jersey.

3. Dr. Basak organized a one-day workshop on *Applied Mathematical Chemistry: Molecular Descriptors and Their Applications in Structure-Property-Activity-Toxicity Relationships*, May 3, 1999, at NRRI. Thirteen speakers from seven different countries, *viz.*, Bulgaria, Croatia, India, Romania, Slovenia, United Kingdom and United States, gave invited presentations on their latest research on Mathematical Chemistry, Quantitative Structure-Activity Relationships (QSAR), Computational Chemistry and Predictive Toxicology.

4. Dr. Basak has been invited to become a member of the International Advisory Committee of the International Symposium *Current Trends in Drug Discovery Research*, February 11-15, 2001, to be organized by the Central Drug Research Institute (CDRI), Lucknow, India, the premier drug discovery and research institute of the country. The symposium is being organized to celebrate the 50th Anniversary of CDRI.

5. Basak has been invited to become a member of the Indian National Organizing Committee of the International Symposium *Strategies and Perspectives in Drug Development, Design and Molecular Modeling* to be organized by the Indian Institute of Chemical Biology, Calcutta, Oct. 17-18, 2000.

6. Dr. S.C. Basak was the Co-Chairperson of the *Second Indo-US Workshop on Mathematical Chemistry with Applications to Drug Discovery, Environmental Toxicology, Cheminformatics and Bioinformatics*, held in Duluth, MN and organized jointly by NRRI and Visva Bharati University, India, May 30-June 3, 2000.

New Discoveries/ Inventions, Patent Disclosures

1. We fond that constituents of complex of mixtures like JP-8 can be clustered into different structural groups using structure spaces derived from topological indices calculated by POLLY

2. An in-depth study of similarity space construction and analog selection resulted in the discovery that for a particular set of compounds the degree of overlap between the groups of analogs selected by theoretical descriptor spaces is relatively high. This study also revealed that a similarity space constructed from physicochemical property data provided relatively unique sets of analogs as compared to those selected from the theoretically-derived similarity spaces.

3. For various sets of toxicological and physicochemical properties the topostructural and topochemical parameters explain most of the variance in the data; the addition of geometrical and quantum chemical parameters to the set of independent variables did small or no improvement in the predicting power of models.

Appendices

Appendix 1 Publications

Appendix 1.1 A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, S.C. Basak, B.D. Gute and G.D. Grunwald, In *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A.T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, p 675-696.

Appendix 1.2 Assessment of the mutagenicity of chemicals from theoretical structural parameters: A hierarchical approach, S.C. Basak, B.D. Gute, and G.D. Grunwald, *SAR QSAR Environ. Res.*, **10**, 117-129, 1999.

Appendix 1.3 Hazard assessment modeling: An evolutionary ensemble approach, D.W. Opitz, S.C. Basak and B.D. Gute, In: *Genetic and Evolutionary Computation*, Eds. W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, & R.E. Smith, Morgan Kaufmann: San Francisco, 1999, p 1643-1651.

Appendix 1.4 Information theoretic indices of neighborhood complexity and their applications, S.C. Basak, In *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A.T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, p 563-593.

Appendix 1.5 Normal boiling points of 1,ω-alkanedinitriles: The highest increment in a homologous series, A.T. Balaban, S.C. Basak and D. Mills, *J. Chem. Inf. Comput. Sci.*, **39**, 769-774, 1999.

Appendix 1.6 A comparative QSAR study of benzamidines complement-inhibitory activity and benzene derivatives acute toxicity, S.C. Basak, B.D. Gute, B. Lucic, S. Nikolic and N. Trinajstic, *Computers & Chemistry*, **24**, 181-191, 2000.

Appendix 1.7 Construction of high-quality structure-property-activity regressions: The boiling points of sulfides, M. Randic and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **40**, 899-905, 2000.

Appendix 1.8 Multiple regression analysis with optimal molecular descriptors, M. Randic and S.C. Basak, SAR QSAR Environ. Res., **11**, 1-23, 2000.

Appendix 1.9 On 3-D graphical representation of DNA primary sequences and their numerical characterization, M. Randic, M. Vracko, A. Nandy and S. C. Basak, *J. Comput. Chem.*, **40**, 1235-1244, 2000.

Appendix 1.10 QSPR modeling: Graph connectivity indices versus line graph connectivity indices, S. C. Basak, S. Nikolic, N. Trinajstic, D. Amic and D. Beslo, *J. Chem. Inf. Comput. Sci.*, **40**, 927-933, 2000.

Appendix 1.11 Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, A. Nandy and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **40**, 915-919, 2000.

Appendix 1.12 Topological indices: Their nature and mutual relatedness, S. C. Basak, A. T. Balaban, G. D. Grunwald and B. D. Gute, *J. Chem. Inf. Comput. Sci.*, **40**, 891-898, 2000.

Appendix 1.13 Use of graph invariants in QMSA and predictive toxicology, S.C. Basak and B.D. Gute, In *Discrete Mathematical Chemistry*, Eds. P. Hansen, P. Fowler, M. Zheng, DIMACS Series 51, American Mathematical Society: Providence, Rhode Island, 2000, p 9-24.

Appendix 1.14 Use of statistical and neural net approaches in predicting toxicity of chemicals, S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian and D. Opitz, *J. Chem. Inf. Comput. Sci.*, **40**, 885-890, 2000.

Appendix 1.1 A hierarchical approach to the development of QSAR models using topological, geometrical...

15. A HIERARCHICAL APPROACH TO THE DEVELOPMENT OF QSAR MODELS USING TOPOLOGICAL, GEOMETRICAL AND QUANTUM CHEMICAL PARAMETERS

S.C. Basak, B.D. Gute and G.D. Grunwald

Natural Resources Research Institute, University of Minnesota-Duluth,
Duluth, MN 55811, USA

A current trend in quantitative structure–property/activity relationship (QSPR/QSAR) studies is the use of theoretical molecular descriptors that can be calculated directly from molecular structure. One advantage of such descriptors is that they can be calculated for any chemical structure, real or hypothetical. Topological indices (TIs) or numerical graph invariants constitute an important subset of these theoretical descriptors. TIs are derived from different classes of weighted graphs, representing various levels of chemical structural information. They are numerical quantifiers of molecular topology and encode information regarding the size, shape, branching pattern, cyclicity, and symmetry of molecular graphs. The Wiener index, different types of connectivity indices, and complexity or information theoretic topological indices have been widely used in QSAR/QSPR research.

We have been involved in the use of TIs in QSAR/QSPR model development to estimate pharmacological, physicochemical, and toxicological properties of diverse sets of molecules. More recently, we have developed a hierarchical approach in the use of theoretical descriptors where topological, geometrical, and quantum chemical indices are used. The goal of this approach has been to use the simplest descriptors first and to only use more complex descriptors if necessary. For this reason, the TIs have been divided into two subsets: (a) topostructural indices (TSIs), the topological indices which are defined on the skeletal molecular graph

In, *Topological Indices and Related Descriptors in QSAR and QSPR* (J. Devillers and A.T. Balaban, Eds.) Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 675–696

and which do not distinguish among the various atoms or bonds present in the molecule, and (b) topochemical indices (TCIs), which explicitly encode information regarding atom and bond types.

In this chapter we will discuss the utility of TIs, geometrical indices, and quantum chemical parameters in hierarchical QSAR studies. The results of studies where the various levels of indices are used in estimating physicochemical, biological, and toxicological properties of different sets of molecules will be presented.

INTRODUCTION

A recent interest in pharmaceutical drug design and hazard assessment of chemicals is the prediction of environmental, physicochemical, toxicological, and pharmacological properties of chemicals directly from their structure [1–11]. Early quantitative structure–activity relationship (QSAR) studies by Hansch and others used physical properties and physicochemical substituent constants for the prediction of other more complex physicochemical, biomedicinal and toxicological properties [12]. Such property–property correlation is useful only when properties necessary for prediction are available for all chemicals under consideration. In the field of environmental risk assessment, most chemicals do not have the data required for proper hazard estimation [13]. In contemporary drug design, one can produce large (real or virtual) combinatorial libraries of chemicals for screening. Most of these chemicals will have no physicochemical data and predictive methods based on experimental data are of no use in this situation. Therefore, there is a need for the development of QSAR methods using nonempirical parameters, *i.e.*, parameters that can be calculated from the molecular structure. Topological indices (TIs), the various molecular size and shape indices as well as quantum chemical parameters fall in this category.

Recently we have developed a new hierarchical approach to QSAR using parameters which are algorithmically defined, *i.e.*, which can be computed from structure using computer software [14–19]. We have successfully used four classes of computed parameters, *viz.*, topostructural, topochemical, geometrical, and quantum chemical parameters, in the development of QSAR models using a hierarchical approach (*vide infra*). This approach was found to be quite useful in the estimation of different properties.

In this chapter we will review the results of our hierarchical QSAR studies pertaining to the prediction of physicochemical, biological, and toxicological properties of different groups of chemicals.

CALCULATION OF PARAMETERS

Computation of Topological Indices

Topological indices used in this study have been calculated by POLLY 2.3 [20] which calculates a total of 102 indices. These indices include the Wiener index [21], the connectivity indices of Kier and Hall [2], and Randić [22], information theoretic indices defined on distance matrices of graphs [23, 24], a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [25–27], and Balaban's *J* indices [28–30]. Table I provides brief definitions for the indices included in this study.

Computation of Geometrical Indices

Van der Waals volume, V_W, [31–33] was calculated using *Sybyl 6.2* [34]. The 3-D Wiener numbers [35] were calculated by *Sybyl* using an SPL (Sybyl Programming Language) program developed in our laboratory. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using *CONCORD 3.2.1* [36]. Two variants of the 3-D Wiener number were calculated: ${}^{3D}W_H$ and ${}^{3D}W$. For ${}^{3D}W_H$, hydrogen atoms are included in the computations and for ${}^{3D}W$, hydrogen atoms are excluded from the computations.

Computation of Quantum Chemical Parameters

The quantum chemical parameters E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔH_f, and μ were calculated for all of the following semi-empirical Hamiltonians: AM1, PM3, MNDO, MINDO/3. These parameters were calculated by *MOPAC 6.00* in the *SYBYL* interface [37]. One difficulty was encountered in using the MINDO/3 Hamiltonian.

Data Reduction and Division of the Topological Indices

Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices and other indices

Table I Symbols, definitions and classifications of topostructural, topochemical, geometrical and quantum chemical descriptors

Topostructural	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC$_r$ reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order $h=0$–6
${}^h\chi_C$	Cluster connectivity index of order $h=3$–6
${}^h\chi_{Ch}$	Chain connectivity index of order $h=3$–6
${}^h\chi_{PC}$	Path-cluster connectivity index of order $h=4$–6
P_h	Number of paths of length $h=0$–10
J	Balaban's J index based on distance
Topochemical	
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC$_r$	Mean information content or complexity of a graph based on the rth ($r=0$–6) order neighborhood of vertices in a hydrogen-filled graph
SIC$_r$	Structural information content for rth ($r=0$–6) order neighborhood of vertices in a hydrogen-filled graph
CIC$_r$	Complementary information content for rth ($r=0$–6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order $h=0$–6
${}^h\chi_C^b$	Bond cluster connectivity index of order $h=3$–6
${}^h\chi_{Ch}^b$	Bond chain connectivity index of order $h=3$–6
${}^h\chi_{PC}^b$	Bond path-cluster connectivity index of order $h=4$–6
${}^h\chi^v$	Valence path connectivity index of order $h=0$–6
${}^h\chi_C^v$	Valence cluster connectivity index of order $h=3$–6
${}^h\chi_{Ch}^v$	Valence chain connectivity index of order $h=3$–6
${}^h\chi_{PC}^v$	Valence path-cluster connectivity index of order $h=4$–6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
Geometrical	
V_W	van der Waals volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix
Quantum Chemical	
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO1}	Energy of the second highest occupied molecular orbital
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO1}	Energy of the second lowest unoccupied molecular orbital
ΔH_f	Heat of formation
μ	Dipole moment

may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of TIs was partitioned into two distinct sets: topostructural indices and topochemical indices. Topostructural indices are indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These sets of the indices are shown in Table I.

To reduce the number of independent variables that were used for model construction in the smaller sets of compounds, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [38]. The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional. From each cluster we select the index most correlated with the cluster, as well as any indices which are poorly correlated with the cluster ($r < 0.70$). These indices are then used in model construction. The variable clustering and selection of indices is performed independently for both the topostructural and topochemical subsets.

DEVELOPMENT OF HIERARCHICAL QSAR MODELS

In the development of hierarchical QSAR models, between two and four sets of indices have been used. A schematic of this method is given in Figure 1 and the SAS procedure REG is used to conduct the all-subsets regression analyses [38]. Final model selection from the all-subsets regression is based on the results for both RSQUARE and CP (Mallow's C_p statistic). The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, the level of complexity is increased one step. To the indices included in the best topostructural model, all of the topochemical indices are added and modeling is conducted using the combined set of parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. Finally, in some studies we have included quantum chemical parameters calculated by MOPAC. The parameters are added to the best model selected from modeling with the combination of topostructural, topochemical and geometrical parameters, and all-subsets regression is used to find the best-fit model. In some of our studies we have also used each level of the hierarchy individually to compare the results of using only one higher-level set, *e.g.*, geometrical indices, alone to determine the degree of contribution to modeling from the given set. Thus, there may be as many as seven final models in a hierarchical study to illustrate the individual contributions of the three higher-level sets of indices, as well as the four models from the stepwise procedure of the hierarchical modeling.

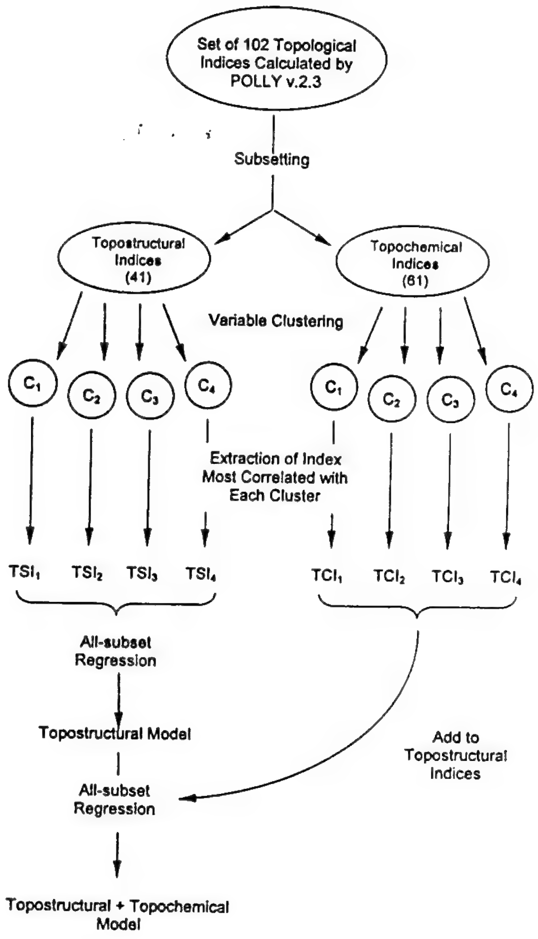

Set of 102 Topological Indices Calculated by POLLY v.2.3
Subsetting
Topostructural Indices (41)
Topochemical Indices (61)
Variable Clustering
C1
C2
C3
C4
C1
C2
C3
C4
Extraction of Index Most Correlated with Each Cluster
TSI1
TSI2
TSI3
TSI4
TCI1
TCI2
TCI3
TCI4
All-subset Regression
Topostructural Model
Add to Topostructural Indices
All-subset Regression
Topostructural + Topochemical Model

Figure 1 Diagramatic representation of the first two stages in hierarchical QSAR model development from topological indices.

HIERARCHICAL QSAR/QSPR STUDIES

The hierarchical method has been used in developing QSAR models for predicting a wide variety of properties. The following are examples from our previous studies employing the hierarchical approach in the construction of useful models.

Physicochemical Properties

Three large sets of chemicals have been used to model physicochemical properties, *viz.*, normal boiling point, lipophilicity ($\log P$), and normal vapor pressure. The normal boiling point data was a subset of the Toxic Substances Control Act (TSCA) Inventory [13] for which measured normal boiling point data were available and where HB_1, a simple measure of the hydrogen bonding potential of a chemical, was equal to zero. This resulted in a set of 1023 diverse chemicals [14]. For this particular set, only the first three levels of the hierarchical approach were used, mainly due to the large amount of computational time necessary to generate quantum chemical parameters for a set of over 1000 chemicals. Eight topostructural indices were selected for the first model (Eq. (1)). The second level of the hierarchy resulted in the retention of two of those topostructural indices and the addition of six topochemical indices (Eq. (2)). Finally, the addition of geometric indices resulted in a ten parameter model using the two topostructural indices, the six topochemical indices, and two of the geometric indices (Eq. (3)). The results of this modeling are presented below (Eqs. (1)–(3)):

$$\begin{aligned} \mathrm{BP} = {} & -21.9 + 30.6(W) - 21.5(O) + 69.9({}^3\chi) + 35.8({}^6\chi) \\ & - 106.5({}^6\chi_C) - 96.1({}^5\chi_{Ch}) - 17.7({}^5\chi_{PC}) + 19.5(P_{10}) \qquad (1) \\ n = 1023, \quad & r^2 = 0.812, \quad s = 39.7°\mathrm{C}, \quad F = 547 \end{aligned}$$

$$\begin{aligned} \mathrm{BP} = {} & -332.9 + 134.6({}^6\chi) + 10.9(P_{10}) + 110.0(\mathrm{IC}_0) - 133.8({}^6\chi^b) \\ & - 80.2({}^3\chi_C^b) + 176.5({}^0\chi^v) + 44.8({}^2\chi^v) + 16.8({}^5\chi_{PC}^v) \qquad (2) \\ n = 1023, \quad & r^2 = 0.961, \quad s = 18.0°\mathrm{C}, \quad F = 3151 \end{aligned}$$

$$\begin{aligned} \mathrm{BP} = {} & -285.7 + 125.3({}^6\chi) + 10.6(P_{10}) + 74.5(\mathrm{IC}_0) - 125.0({}^6\chi^b) \\ & - 86.3({}^3\chi_C^b) + 175.3({}^0\chi^v) + 49.1({}^2\chi^v) + 18.7({}^5\chi_{PC}^v) \\ & - 9.1({}^{3D}W_H) + 8.1({}^{3D}W) \qquad (3) \\ n = 1023, \quad & r^2 = 0.963, \quad s = 17.6°\mathrm{C}, \quad F = 2650. \end{aligned}$$

From the three equations presented, it is clear that the replacement of six topostructural indices with six topochemical indices greatly enhanced the predictive power of the model, while the addition of the geometric parameters did not add much to the model. A scatterplot of experimental versus predicted boiling point from Eq. (3) is shown in Figure 2.

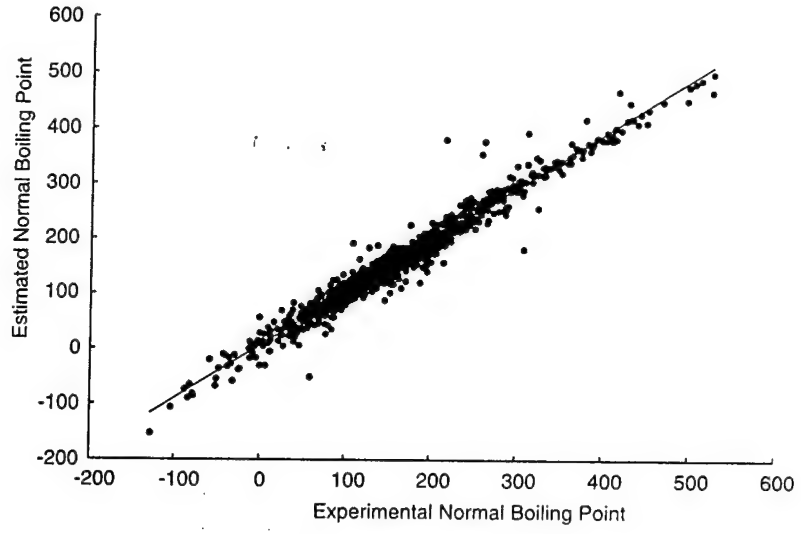


Figure 2 Scatterplot of experimental normal boiling point *vs* estimated normal boiling point using Eq. (3) for 1023 diverse chemicals.

The lipophilicity data are a subset of 219 chemicals derived from the STARLIST set with $\log P$ values between −2 to 5.5 obtained from CLOGP [39] and HB_1 equal to zero [14]. This subset was chosen to examine the effectiveness of model based on topological indices in the prediction of lipophilicity for compounds that do not have explicit hydrogen-bonding centers. Compounds were chosen within the range of $\log P$ values described to avoid the problematic nature of compounds having exceptionally high values for lipophilicity. As with the boiling point models, only the first three levels of the hierarchy were applied to modeling lipophilicity. Seven topostructural indices were initially selected (Eq. (4)), and again, only two were retained with the addition of eight topochemical indices (Eq. (5)). In Eq. (6), with the addition of two geometric parameters, an additional topostructural index is removed from the model. These equations are presented below:

$$\begin{aligned} \log P = {} & -1.42 + 1.08(W) - 1.58({}^2\chi) + 1.51({}^6\chi) - 0.92({}^6\chi_C) \\ & - 0.32(P_7) + 0.20(P_{10}) + 1.97(J) \qquad (4) \\ n = 219, \quad & r^2 = 0.789, \quad s = 0.54, \quad F = 112 \end{aligned}$$

$$\log P = -2.13 - 0.20(^2\chi) + 0.18(P_{10}) - 1.86(\mathrm{IC}_0) + 1.33(\mathrm{CIC}_2) - 0.92(\mathrm{CIC}_3) - 1.36(^6\chi^b) + 5.76(^0\chi^v) - 2.98(^1\chi^v) + 0.54(^4\chi^v) - 0.39(^3\chi_C^v) \quad (5)$$
$$n = 219, \quad r^2 = 0.908, \quad s = 0.36, \quad F = 206$$

$$\log P = -5.60 + 0.19(P_{10}) - 1.46(\mathrm{IC}_0) + 1.09(\mathrm{CIC}_2) - 0.77(\mathrm{CIC}_3) - 1.36(^6\chi^b) + 5.34(^0\chi^v) - 3.41(^1\chi^v) + 0.55(^4\chi^v) - 0.41(^3\chi_C^v) + 1.10(V_W) - 0.17(^{3D}W) \quad (6)$$
$$n = 219, \quad r^2 = 0.912, \quad s = 0.35, \quad F = 194.$$

These three equations show similar results as those for the modeling of normal boiling point. The replacement of topostructural indices with an equal or greater number of topochemical indices results in marked improvement in the predictive power of the model, while the addition of geometric indices resulted in only a minor improvement. Figure 3 presents a plot of the experimental log P values versus the log P values predicted from Eq. (6). The 219 chemicals and their observed and predicted values for log P have been presented previously in the literature [14].

The 476 chemicals in the normal vapor pressure data [16] are a subset of the TSCA inventory taken from the ASTER (Assessment Tools for the Evaluation of Risk) database [40]. This is a diverse subset of chemicals all have vapor pressure (p_{vap}) data measured at 25°C and ranging between 3–10,000 mmHg.

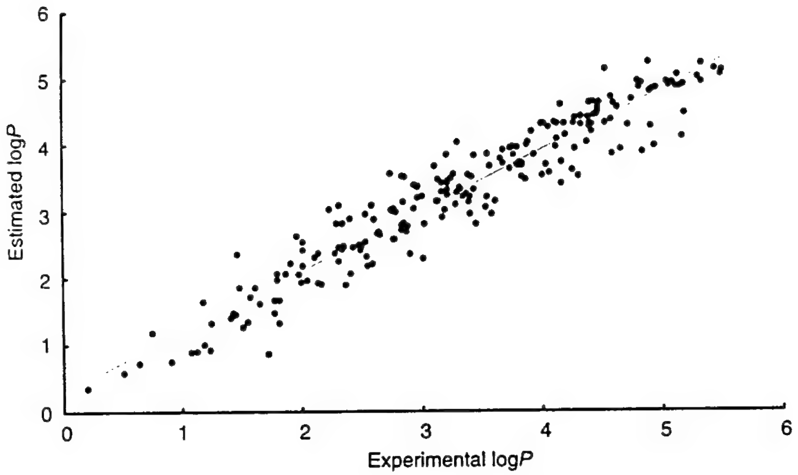


Figure 3 Scatterplot of experimental log P vs estimated log P using Eq. (6) for 219 diverse chemicals.

The first three levels of the hierarchical method have been employed; however, the addition of geometric parameters to the modeling process did not result in the selection of a novel model and so there is no geometric model reported.

$$\log_{10}(p_{\mathrm{vap}}) = 4.88 + 0.20(O) - 2.56(^1\chi) + 0.49(^4\chi_C) + 0.79(^6\chi_C) + 0.98(P_{10}) \quad (7)$$
$$n = 476, \quad r^2 = 0.515, \quad s = 0.53, \quad F = 99.7$$

$$\log_{10}(p_{\mathrm{vap}}) = 8.44 - 1.77(^1\chi) + 1.25(P_{10}) - 5.69(\mathrm{IC}_1) + 3.91(\mathrm{IC}_2) - 1.24(\mathrm{IC}_5) + 1.41(^3\chi_C^b) - 1.70(^1\chi^v) \quad (8)$$
$$n = 476, \quad r^2 = 0.793, \quad s = 0.34, \quad F = 224.0.$$

As can be seen from Eq. (7), five topostructural indices were initially selected to model normal vapor pressure. The addition of the topochemical indices resulted in the retention of two topostructural indices and the addition of five topochemical indices (Eq. (8)). As was seen for the other two physicochemical properties, *viz.*, normal boiling point and lipophilicity, the predictive power of the model is greatly enhanced by the addition of the topochemical indices. A scatterplot of experimental versus predicted normal vapor pressure, based on Eq. (8), is shown in Figure 4. These results are adequate, however, as can be seen from Figure 5 while the residuals show fairly uniform scatter when plotted against the dependent variable there are some significant outliers and the data tends to be somewhat skewed to the lower end of the vapor pressure range.

Biological Properties

Two smaller sets of congeneric chemicals have been used in the study of biological properties. The smaller of the two sets [19] consisted of sixty polycyclic aromatic hydrocarbons for which 24-hour dermal penetration (DP) data were available from the work of Roy *et al.* [41]. For the purposes of this study, all four levels of the hierarchical method were employed. Only two equations are being presented since the addition of geometric and quantum chemical parameters to the modeling procedure did not result in the formulation of improved QSAR equations.

$$\mathrm{DP} = 224.1 - 67.9(P_0) \quad (9)$$
$$n = 60, \quad r^2 = 0.675, \quad s = 7.4, \quad F = 120.6$$

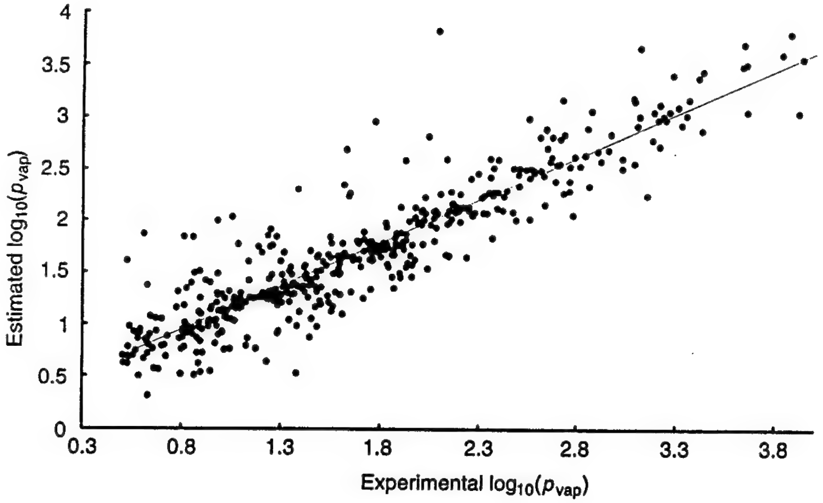


Figure 4 Scatterplot of experimental normal vapor pressure *vs* estimated normal vapor pressure using Eq. (8) for 476 diverse chemicals.

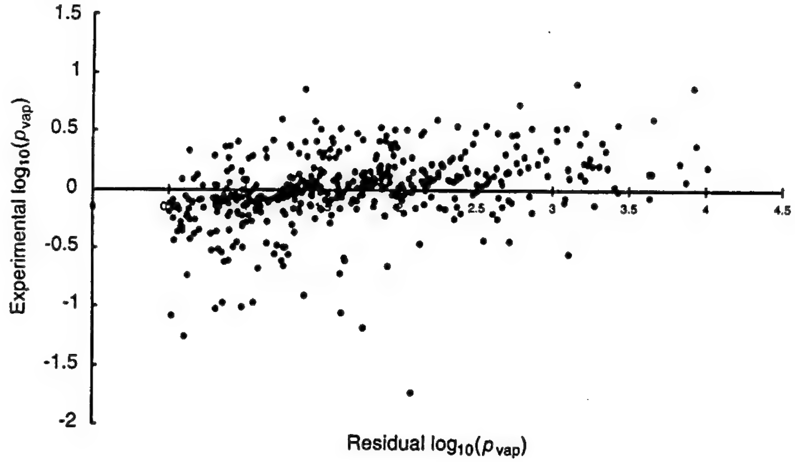


Figure 5 Scatterplot of the residual *vs* experimental normal vapor pressure from Eq. (8) for 476 diverse chemicals.

$$\mathrm{DP} = 179.7 - 78.8({}^{1}\chi^{b}) \qquad (10)$$
$$n = 60, \quad r^2 = 0.695, \quad s = 7.1, \quad F = 132.0.$$

Eq. (9) shows the model resulting from the topostructural modeling. A one parameter model which explains 67.5% of the variance was generated. A small improvement is seen in the model resulting from the addition of the topochemical indices (Eq. (10)), in which the topostructural index is replaced by the topochemical index, ${}^{1}\chi^{b}$. Figure 6 presents a scatterplot of experimental dermal penetration versus the predicted results from Eq. (10).

The second set of biological data studied using the hierarchical method was a set of 107 benzamidines [18] that act as inhibitors of the complement system, collected from the literature by Hansch and Yoshimoto [42]. The base structure for the benzamidines is presented in Figure 7 and the side-chains and activity values have been published previously [18]. The large size of these molecules made the calculation of quantum chemical indices prohibitively time-consuming. As a result, the first three levels of the hierarchical modeling procedure were used for this study.

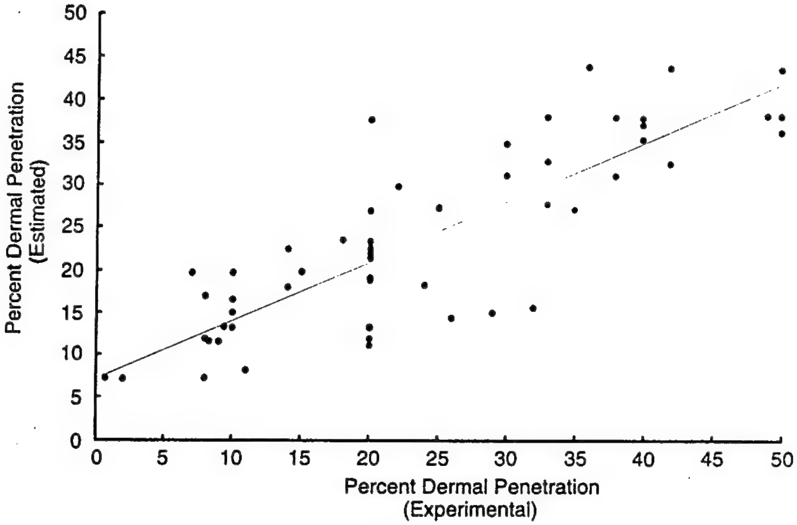


Figure 6 Scatterplot of experimental percent dermal penetration *vs* estimated percent dermal penetration using Eq. (10) for 60 polycyclic aromatic hydrocarbons.

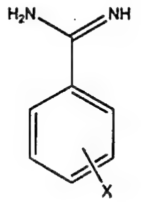


Figure 7 Neutral base structure for the 107 benzamidines.

$$1/\log C = 1.1245 + 0.4989(I^{D}) \qquad (11)$$
$$n = 105, \quad r^2 = 0.884, \quad s = 0.0200, \quad F = 785$$

$$1/\log C = -0.6428 + 0.0490(^{3D}W) \qquad (12)$$
$$n = 105, \quad r^2 = 0.889, \quad s = 0.0196, \quad F = 824.$$

A single topostructural index provided a strong correlation with the inhibitory activity of these large compounds (Eq. (11)). This one index modeled the activity so well, that the addition of topochemical indices did not add significantly to the predictive power of the model. Finally, with the addition of geometric parameters to the modeling of inhibitory activity, it was found that one geometric parameter provided a slightly better correlation with activity than did the topostructural index (Eq. (12)), explaining 89% of the variance in the data. The results of this final model (Eq. (12)) are shown in Figure 8 as a scatterplot of experimental versus predicted activity.

Toxicological Properties

Two sets of compounds have been studied using the hierarchical modeling for toxicological properties. The first set consists of acute aquatic toxicity data for 69 benzene derivatives determined by the 96-hour fathead minnow toxicity test system [17]. This data was compiled by Hall, Kier, and Phipps [43] from eight literature sources and was supplemented by some original work conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota.

$$LC_{50} = -7.50 + 3.50(M_1) - 1.72(\overline{IC}) - 0.52(P_8) + 0.68(P_9) \qquad (13)$$
$$n = 69, \quad r^2 = 0.453, \quad s = 0.58, \quad F = 13.3$$

$$LC_{50} = 23.68 + 5.04(M_1) + 0.55(P_9) - 43.27(SIC_0) - 20.04(CIC_0) \quad (14)$$
$$n = 69, \quad r^2 = 0.783, \quad s = 0.36, \quad F = 57.9$$

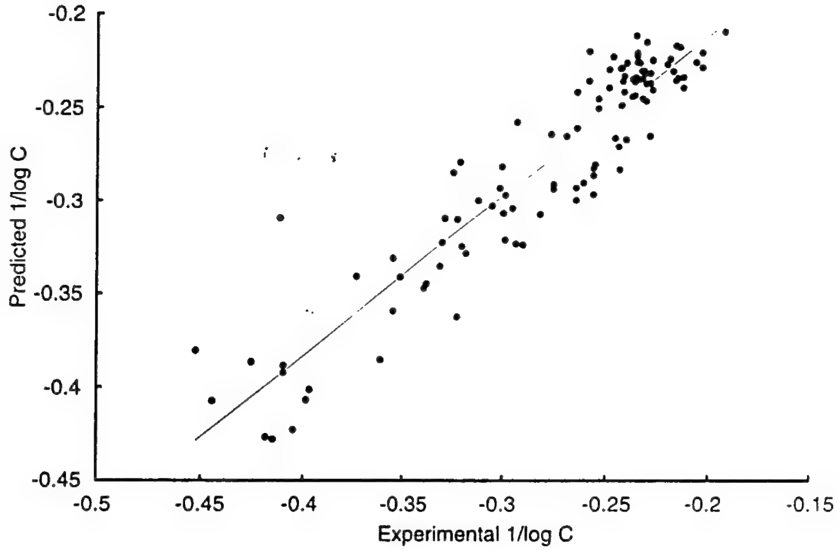


Figure 8 Scatterplot of experimental complement inhibition *vs* estimated complement inhibition using Eq. (12) for 105 benzamidines.

$$LC_{50} = 0.59 + 5.82(M_1) + 0.55(P_9) - 14.23(SIC_0) - 2.36(^{3D}W_H) \quad (15)$$
$$n = 69, \quad r^2 = 0.792, \quad s = 0.36, \quad F = 61.1$$

$$LC_{50} = -3.83 + 5.97(M_1) + 0.77(P_9) - 8.26(SIC_0) - 1.98(^{3D}W_H) + 0.41(E_{LUMO1}) + 0.01(\Delta H_f) - 0.12(\mu) \quad (16)$$
$$n = 69, \quad r^2 = 0.863, \quad s = 0.30, \quad F = 55.0.$$

Eq. (13) shows the results of the initial modeling using topostructural indices. Even using four indices, the topostructural set did a poor job of modeling acute toxicity. The addition of topochemical indices led to a significant improvement in predictive power, with the replacement of two topostructural indices with topochemical indices (Eq. (14)). The geometrical indices slightly improved the QSAR modeling (Eq. (15)); however, it was the addition of quantum chemical indices which drastically improved the predictive power of our model (Eq. (16)). The addition of quantum chemical indices increased the variance explained by 7.1% over

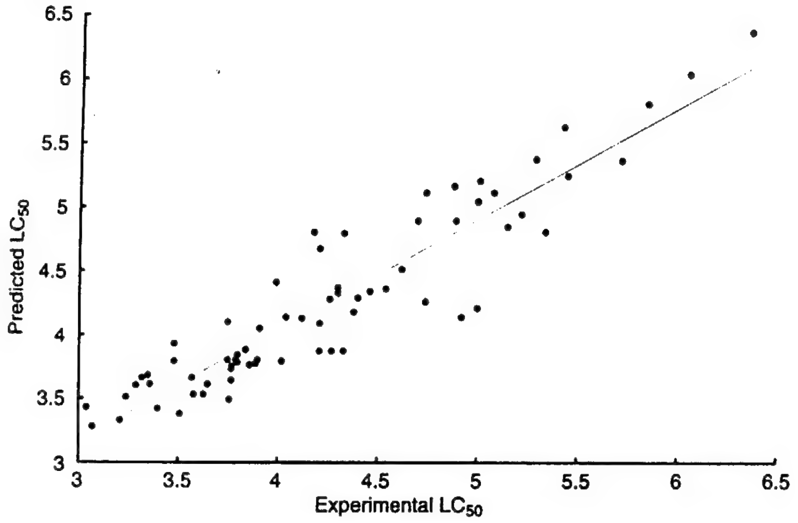


Figure 9 Scatterplot of experimental acute aquatic toxicity (LC_{50}) vs estimated acute aquatic toxicity using Eq. (16) for 69 benzene derivatives.

the model including geometrical indices, resulting in an overall explanation of 86.3% of the variance. Figure 9 presents the scatterplot of experimental versus predicted toxicity for these 69 compounds based on the results of Eq. (16).

A set of 520 compounds, 260 mutagens and 260 non-mutagens, was taken from the literature [44] as a source of mutagenicity data. These data provided qualitative assessments of mutagenicity based on a positive or negative result in the Ames' mutagenicity assay. A discriminant function analysis (DFA) was conducted on this set using the SAS procedure DISCRIM [38] to create a function capable of classifying the compounds as active or inactive. Based on the results of a previous study and the amount of time required for the calculations, the quantum chemical parameters were excluded and indicators of molecular fragments associated with mutagenic activity were included [15]. See the original manuscript for a further discussion of the data used in this study and the molecular fragments keyed for the analysis. These classification results, the indices used in each case, and brief notes on the fragment groups included in the final models are presented in Table II.

Table II Classification results for 520 mutagens/non-mutagens from DFA

Model Type	*Indices Included*	*% Mutagens Correct*	*% Non-mutagens Correct*
Topostructural	W, H^V, H^D, IC, M_1, ${}^2\chi$, ${}^3\chi$, ${}^4\chi$, ${}^6\chi_C$, ${}^6\chi_{PC}$, P_{10}	76.2	57.3
Topostructural + topochemical	H^D, M_1, ${}^2\chi$, P_{10}, IC_5, ${}^6\chi^b_{Ch}$, ${}^0\chi^v$, ${}^2\chi^v$, ${}^3\chi^v_{Ch}$, ${}^6\chi^v_{Ch}$, ${}^6\chi^v_{PC}$, J^X, J^B	74.6	63.1
Topostructural + topochemical + fragments	H^D, M_1, ${}^2\chi$, P_{10}, IC_5, ${}^0\chi^v$, ${}^3\chi^v_{Ch}$, ${}^6\chi^v_{PC}$, J^B, nitroso[1], mustard[2], sulf[3], benz[4]	69.2	71.9
Topostructural + topochemical + fragments + geometrical	H^D, M_1, ${}^2\chi$, P_{10}, IC_5, ${}^0\chi^v$, ${}^3\chi^v_{Ch}$; ${}^6\chi^v_{PC}$, J^B, nitroso[1], mustard[2], sulf[3], benz[4], V_W	71.5	71.9

[1] Nitroso-compounds. [2] Halogenated substituted mustard, sulfur mustard or oxygen mustard. [3] Organic sulfates or sulfonates. [4] Biphenyl amine, benzidine or 4,4′-methylenedianiline derivatives.

As can be seen in Table II, the topostructural indices alone correctly classify over 75% of the mutagens; however, they only correctly classify 57.3% of the non-mutagens. This leaves over 40% of the non-mutagens incorrectly classified. The combination of topostructural and topochemical indices results in a comparable classification rate for mutagens (74.6%) and a significant increase (5.8%) in the classification of non-mutagens. The addition of information regarding the presence or absence of known structural fragments associated with mutagenic activity results in a significant decrease (5.4%) in classification rate for mutagens, from 74.6% down to 69.2%. However, the addition of these structural fragments also increases the correct classification rate for non-mutagens increasing it from 63.1% to 71.9%, and overall increase of 8.7%. As a result of this dramatic increase in classification rate for non-mutagens, this model was retained and supplemented by the geometrical indices. Addition of the geometric indices brought the classification rate for mutagens up to 71.5% (an overall decrease of 4.7% from the topostructural model) and retained the classification rate for non-mutagens at 71.9% (an overall increase of 14.6% over the initial model). While these results are by no means spectacular, it is a reasonably accurate model for the prediction of mutagenic activity.

CONCLUSION

The goal of hierarchical QSAR studies is to investigate the relative roles of different classes of parameters, *viz.*, topostructural and topochemical indices, 3-D parameters and calculated quantum chemical parameters in predicting different types of molecular properties. It is clear from the results presented here that topostructural and topochemical indices explain most of the variance in the data for physicochemical, biological and toxicological properties. In most cases geometrical and quantum chemical indices make only marginal improvements in the predictive power of the models. This indicates that the easily calculable topostructural and topochemical indices will be an effective first choice in QSAR studies.

It is evident from these studies that the expanded levels of the hierarchical method are extremely useful for large, diverse sets of chemicals where there are many factors influencing the variation of properties between chemical structures. They are also useful in modeling the more complex biological interactions involving the modulation of toxicants.

It is interesting to note that studies involving the inhibition of a specific enzymatic system or the passage of large compounds through the skin are modeled well using simply shape and size descriptors, and do not seem to benefit significantly from the addition of more complex indices. There is still a need for better descriptors that will help us to more accurately model complex biological and toxicological systems.

Acknowledgements

This paper is contribution number 236 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force.

References

[1] Randić, M. (1984). Nonempirical approaches to structure–activity studies. *Int. J. Quantum Chem: Quant. Biol. Symp.* **11**, 137–153.

[2] Kier, L.B. and Hall, L.H. (1986). *Molecular Connectivity in Structure–Activity Analysis*. Research Studies Press: Letchworth, Hertfordshire, U.K, p. 262.

[3] Rouvray, D.H. and Pandey, R.B. (1986). The fractal nature, graph invariants and physicochemical properties of normal alkanes. *J. Chem. Phys.* **85**, 2286–2290.

[4] Basak, S.C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR Approach. *Med. Sci. Res.* **15**, 605–609.

[5] Basak, S.C., Frane, C.M., Rosen, M.E., and Magnuson, V.R. (1987). Molecular topology and acute toxicity: A QSAR study of monoketones. *Med. Sci. Res.* **15**, 887–888.

[6] Basak, S.C. (1988). Binding of barbiturates to cytochrome P_{450}: A QSAR study using log P and topological indices. *Med. Sci. Res.* **16**, 281–282.

[7] Basak, S.C. (1990). A nonempirical approach to predicting molecular properties using graph-theoretic invariants. In, *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, Eds.). Kluwer Academic Publishers, Dordrecht, pp. 83–103.

[8] Basak, S.C., Niemi, G.J., and Veith, G.D. (1991). Predicting properties of molecules using graph invariants. *J. Math. Chem.* **7,** 243–272.

[9] Balaban, A.T., Basak, S.C., Colburn, T., and Grunwald, G. (1994). Correlation between structure and normal boiling points of haloalkanes C_1–C_4 using neural networks. *J. Chem. Inf. Comput. Sci.* **34,** 1118–1121.

[10] Basak, S.C. and Grunwald, G.D. (1995). Predicting genotoxicity of chemicals using nonempirical parameters. In, *Proceeding of the XVI International Cancer Congress* (R.S. Rao, M.G. Deo, and L.D. Sanghui, Eds.). Monduzzi Bologna, Italy, pp. 413–416.

[11] Basak, S.C., Grunwald, G.D., and Niemi, G.J. (1997). Use of graph-theoretic and geometrical molecular descriptors in structure–activity relationships. In, *From Chemical Topology to Three-Dimensional Geometry* (A.T. Balaban, Ed.). Plenum Press, New York, pp. 73–116.

[12] Hansch, C. and Leo, A. (1995). *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology.* American Chemical Society, Washington, D.C., p. 557.

[13] Auer, C.M., Nabholz, J.V., and Baetcke, K.P. (1990). Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure–activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **87,** 183–197.

[14] Basak, S.C., Gute, B.D., and Grunwald, G.D. (1996). A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **36,** 1054–1060.

[15] Basak, S.C. and Grunwald, G.D. (1995). Predicting genotoxicity of chemicals using nonempirical parameters. In, *Proceedings of the XVI International Cancer Congress* (R.S. Rao, M.G. Deo, and L.D. Sanghui, Eds.). Monduzzi, Bologna, Italy, Vol. 7, pp 413–416.

[16] Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **37,** 651–655.

[17] Gute, B.D. and Basak, S.C. (1997). Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **7,** 117–131.

[18] Basak, S.C., Gute, B.D., and Grunwald, G.D. (1999). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.* **39,** 255–260.

[19] Gute, B.D., Grunwald, G.D., and Basak, S.C. (1999). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **10,** 1–15.

[20] Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1988). POLLY 2.3: Copyright of the University of Minnesota.

[21] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69,** 17–20.

[22] Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.* **97,** 6609–6615.

[23] Raychaudhury, C., Ray, S.K., Ghosh, J.J., Roy, A.B., and Basak, S.C. (1984). Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.* **5,** 581–588.

[24] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **67,** 4517–4533.

[25] Basak, S.C., Roy, A.B., and Ghosh, J.J. (1980). Study of the structure–function relationship of pharmacological and toxicological agents using information theory. In, *Proceedings of the Second International Conference on Mathematical Modelling* (X.J.R. Avula, R. Bellman, Y.L. Luke, and A.K. Rigler, Eds.). University of Missouri—Rolla, pp. 851–856.

[26] Basak, S.C. and Magnuson, V.R. (1983). Molecular topology and narcosis: A quantitative structure–activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim. Forsch.* **33,** 501–503.

[27] Roy, A.B., Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications. In, *Mathematical Modelling in Science and Technology* (X.J.R. Avula, R.E. Kalman, A.I. Lapis, and E.Y. Rodin, Eds.). Pergamon Press, New York, pp. 745–750.

[28] Balaban, A.T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89,** 399–404.

[29] Balaban, A.T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* **55,** 199–206.

[30] Balaban, A.T. (1986). Chemical graphs. Part 48. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem.* (*MATCH*). **21,** 115–122.

[31] Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.* **68,** 441–451.

[32] Moriguchi, I., Kanada, Y., and Komatsu, K. (1976). Van der Waals volume and the related parameters for hydrophobicity in structure–activity studies. *Chem. Pharm. Bull.* **24,** 1799–1806.

[33] Moriguchi, I., and Kanada, Y. (1977). Use of van der Waals volume in structure–activity studies. *Chem. Pharm. Bull.* **25,** 926–935.

[34] *SYBYL Version 6.1.* (1994). Tripos Associates, Inc.: St. Louis, MO.

[35] Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstić, N., and Bangov, I. (1986). Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. *Arzneim.-Forsch./Drug Research* **36,** 176–183.

[36] *CONCORD Version 3.0.1.* (1993). Tripos Associates, Inc.: St. Louis, MO.

[37] Stewart, J.J.P. (1990). MOPAC Version 6.00. QCPE #455. Frank J. Seiler Research Laboratory: US Air Force Academy, CO.

[38] SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition.* SAS Institute Inc.: Cary, NC.

[39] Leo, A. and Weininger, D. (1984). *CLOGP Version 3.2 User Reference Manual.* Medicinal Chemistry Project, Pomona College, Claremont, CA.

[40] Russom, C.L., Anderson, E.B., Greenwood, B.E., and Pilli, A. (1991). ASTER: An integration of the AQUIRE data base and the QSAR system for use in ecological risk assessments. *Sci. Total Environ.* **109/110,** 667–670.

[41] Roy, T.A., Neil, W., Yang, J.J., Krueger, A.J., Arroyo, A.M., and Mackerer, C.R. (1998). SAR models for estimating the percutaneous absorption of polynuclear aromatic hydrocarbons. *SAR QSAR Environ. Res.* **9,** 171–185.

[42] Hansch, C. and Yoshimoto, M. (1974). Structure–activity relationships in immunochemistry. 2. Inhibition of complement by benzamidines. *J. Med. Chem.* **17,** 1160–1167.

[43] Hall, L.H., Kier, L.B., and Phipps, G. (1984). Structure–activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.* **3,** 355–365.

[44] Soderman, J.V. (1982). *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity–Mutagenicity Database*, CRC Press, Inc., Boca Raton, FL, Vol. I, p. 655.

16. MOLECULAR GRAPH DESCRIPTORS USED IN NEURAL NETWORK MODELS

O. Ivanciuc

Department of Organic Chemistry, Faculty of Chemical Technology
University "Politehnica" of Bucharest, Oficiul 12 CP 243,
78100 Bucharest, Romania

Artificial neural networks are general nonlinear models that proved to be very efficient in computing physical, chemical, and biological properties of various classes of chemical compounds. The success of neural networks in structure–property models depends mainly on the numerical representation of the structure of the compounds in network calibration and prediction. The atomic and molecular graph descriptors used as input data for neural networks are presented together with examples of their computation. Three new neural networks were defined in order to encode into their topology the chemical structure of each compound presented to the network: the Baskin–Palyulin–Zefirov neural device, ChemNet defined by Kireev, and MolNet introduced by Ivanciuc. All three neural models use information from the molecular graph to generate the neural network. The rules that define the three neural models are presented together with examples of network generation from the molecular graph.

INTRODUCTION

Conventional quantitative structure–property relationship (QSPR) and quantitative structure–activity relationship (QSAR) models require the user to specify the mathematical function of the model. If these functions

In, *Topological Indices and Related Descriptors in QSAR and QSPR* (J. Devillers and A.T. Balaban, Eds.) Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 697–777

APPENDIX 1.2 Assessment of the mutagenicity of chemicals from theoretical structural parameters

SAR and QSAR in Environmental Research
1999, Vol. 10, pp. 117–129
Reprints available directly from the publisher
Photocopying permitted by license only

ASSESSMENT OF THE MUTAGENICITY OF AROMATIC AMINES FROM THEORETICAL STRUCTURAL PARAMETERS: A HIERARCHICAL APPROACH*

S. C. BASAK[†], B. D. GUTE and G. D. GRUNWALD

*Natural Resources Research Institute, 5013 Miller Trunk Hwy.,
Duluth, MN 55811, USA*

(Received 8 July 1998; In final form 16 October 1998)

A hierarchical approach has been used in this paper in predicting the mutagenicity/non-mutagenicity of a set of 127 chemicals from their molecular descriptors. The set of descriptors consisted of topostructural and topochemical parameters, experimental properties like log P, and quantum chemical indices calculated using a semi-empirical method. The results show that a combination of topostructural and topochemical molecular descriptors explain most of the variance in the experimental data. The addition of physical properties or quantum chemical parameters did not make any significant improvement in the predictive power of the models.

Keywords: Aromatic amines; hierarchical similarity; mutagenicity; quantum chemical descriptors; topological indices

INTRODUCTION

A current interest in the fields of chemistry, toxicology and biomedical sciences is the prediction of the property/activity of chemicals from calculated molecular descriptors [1–6]. In both environmental hazard assessment and pharmaceutical drug design, one has to deal with thousands, sometimes millions, of real or hypothetical chemical structures. Most of these compounds have very little of the experimental data necessary for the

*Presented at the workshop *Computational Methods in Toxicology*, April 20–22, 1998, Dayton, OH, USA.

[†]Corresponding author.

estimation of their toxicity or efficacy. In this age of combinatorial chemistry, one can synthesize thousands of chemicals very quickly. However, experimental testing of these large numbers of chemicals would not be cost effective. Also, it is possible to create virtual libraries consisting of billions of structures. In this case one would like to know the toxic, as well as therapeutic, potential of such a vast collection of chemicals. The experimental data necessary for the prediction of the toxicity/activity of these large and diverse sets of chemicals will not be available to us in the near future.

This pervasive lack of experimental data demonstrates the need for the development of predictive models based on parameters that can be calculated directly from a chemical's molecular structure. Recently, our research group has been involved in the development of a hierarchical approach to quantitative structure-activity relationship (QSAR) model development for predicting physicochemical, toxicological and pharmacological properties of chemicals using theoretical molecular descriptors [3, 6 – 10]. Various topological indices (TIs) fall in this category of molecular descriptors [11 – 23]. Balaban has classified TIs into three generations based on whether they are integers, real numbers or a sequence of numbers [24]. Different classes of TIs quantify various aspects of molecular structure. We have shown in the past that various indices, *viz.*, connectivity indices and complexity indices developed and used by Basak *et al.* [15 – 18] quantify distinctly different types of molecular structural information. Such indices can be calculated very rapidly. On the other hand, geometrical and quantum chemical parameters encode information regarding the stereo-electronic aspects of molecules. These classes of parameters are also algorithmically derived, *i.e.*, they can be calculated for any real or hypothetical molecular structure without any input of experimental data.

One of our recent interests has been to test the relative effectiveness of the four classes of theoretical molecular descriptors mentioned above in the development of QSARs for predicting property/activity/toxicity of chemicals [3, 6 – 10]. In this paper we have used these parameters in the development of models for predicting mutagenicity/non-mutagenicity of a set of 127 aromatic amines.

METHODS

Datasets

A set of 127 aromatic and heteroaromatic amines, previously collected from the literature by Debnath *et al.* [25], were used to study mutagenicity. The mutagenicity of these compounds in *S. Typhimurium* TA98 + S9 microsomal preparation has been expressed as positive or negative mutagenicity by Benigni [26]. Compounds included in this study and their mutagenic classification based on experimentally determined mutagenic potency are given in Table I. Of the compounds used in this study, 106 were classified as mutagens while twenty-one were determined to be non-mutagens.

TABLE I Aromatic and heteroaromatic amines[1]

Chemicals	*TA*98 (*Expt.*)	*TA*98 (*Pred.*)[2]
2-Bromo-7-aminofluorene	1	1
2-Methoxy-5-methylaniline (*p*-cresidine)	1	1
5-Aminoquinoline	1	1
4-Ethoxyaniline (*p*-phenetidine)	1	1
1-Aminonaphthalene	1	1
4-Aminofluorene	1	1
2-Aminoanthracene	1	1
7-Aminofluoranthene	1	1
8-Aminoquinoline	1	1
1,7-Diaminophenazine	1	1
2-Aminonaphthalene	1	1
4-Aminopyrene	1	1
3-Amino-3′-nitrobiphenyl	1	1
2,4,5-Trimethylaniline	1	1
3-Aminofluorene	1	1
3,3′-Dichlorobenzidine	1	1
2,4-Dimethylaniline (2,4-xylidine)	1	1
2,7-Diaminofluorene	1	1
3-Aminofluoranthene	1	1
2-Aminofluorene	1	1
2-Amino-4′-nitrobiphenyl	1	1
4-Aminobiphenyl	1	1
3-Methoxy-4-methylaniline (*o*-cresidine)	1	0
2-Aminocarbazole	1	1
2-Amino-5-nitrophenol	1	1
2,2′-Diaminobiphenyl	1	1
2-Hydroxy-7-aminofluorene	1	1
1-Aminophenanthrene	1	1
2,5-Dimethylaniline (2,5-xylidine)	1	1
4-Amino-2′-nitrobiphenyl	1	1
2-Amino-4-methylphenol	1	1
2-Aminophenazine	1	1
4-Aminophenylsulfide	1	1
2,4-Dinitroaniline	1	1
2,4-Diaminoisopropylbenzene	1	1
2,4-Difluoroaniline	1	1
4,4′-Methylenedianiline	1	1
3,3′-Dimethylbenzidine	1	1
2-Aminofluoranthene	1	1
2-Amino-3′-nitrobiphenyl	1	1
1-Aminofluoranthene	1	1

TABLE I (Continued)

Chemicals	*TA*98 (*Expt.*)	*TA*98 (*Pred.*)[2]
4,4′-Ethylenebis (aniline)	1	1
4-Chloroaniline	1	1
2-Aminophenanthrene	1	1
4-Fluoroaniline	1	1
9-Aminophenanthrene	1	1
3,3′-Diaminobiphenyl	1	1
2-Aminopyrene	1	1
2,6-Dichloro-1,4-phenylenediamine	1	1
2-Amino-7-acetamidofluorene	1	1
2,8-Diaminophenazine	1	1
6-Aminoquinoline	1	1
4-Methoxy-2-methylaniline (*m*-cresidine)	1	1
3-Amino-2′-nitrobiphenyl	1	1
2,4′-Diamino-biphenyl	1	1
1,6-Diaminophenazine	1	1
4-Aminophenyldisulfide	1	1
2-Bromo-4,6-dinitroaniline	1	1
2,4-Diamino-*n*-butylbenzene	1	0
4-Aminophenylether	1	1
2-Aminobiphenyl	1	1
1,9-Diaminophenazine	1	1
1-Aminofluorene	1	1
8-Aminofluoranthene	1	1
2-Chloroaniline	1	0
2-Amino-aaa-trifluorotoluene	1	1
2-Amino-1-nitronaphthalene	1	1
3-Amino-4′-nitrobiphenyl	1	1
4-Bromoaniline	1	1
2-Amino-4-chlorophenol	1	1
3,3′-Dimethoxybenzidine	1	1
4-Cyclohexylaniline	1	1
4-Phenoxyaniline	1	1
4,4′-Methylenebis (*o*-ethylaniline)	1	0
2-Amino-7-Nitrofluorene	1	1
Benzidine	1	1
1-Amino-4-Nitronaphthalene	1	1
4-Amino-3′-Nitrobiphenyl	1	1
4-Amino-4′-Nitrobiphenyl	1	1
1-Aminophenazine	1	1
4,4′-Methylenebis (*o*-fluoroaniline)	1	1
4-Chloro-2-nitroaniline	1	1
3-Aminoquinoline	1	1
3-Aminocarbazole	1	1
4-Chloro-1,2-phenylenediamine	1	1
3-Aminophenanthrene	1	1
3,4′-Diaminobiphenyl	1	1
1-Aminoanthracene	1	1
1-Aminocarbazole	1	1
9-Aminoanthracene	1	1
4-Aminocarbazole	1	1
6-Aminochrysene	1	1
1-Aminopyrene	1	1
4-4′-Methylenebis (*o*-isopropyl-aniline)	1	0

TABLE I (Continued)

Chemicals	*TA98* *(Expt.)*	*TA98* *(Pred.)*[2]
2,7-Diaminophenazine	1	1
4-Aminophenanthrene	0	1
2,4-Diaminotoluene	1	1
3,3′-Diaminobenzidine	1	1
1,3-Phenylenediamine	1	0
3,4-Diaminotoluene	1	1
1,2-Phenylenediamine	1	0
3-Amino-6-methylphenol	1	1
2,4-Diaminoethylbenzene	1	1
4,4′-Methylenebis (2,6-diisopropylaniline)	0	0
4,4′-Methylenebis (2,6-diethylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-*t*-butylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-isopropylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-ethylaniline)	0	0
4,4′-Methylenebis (2,6-dimethylaniline)	0	1
3-Aminobiphenyl	0	1
2,3-Diaminobiphenyl	0	1
2-Methyl-4-chloroaniline	0	1
2-Chloro-4-methylaniline	0	1
4-Methoxyaniline	0	1
3-Methoxyaniline	0	1
Aniline	0	0
3-Chloroaniline	0	0
3-Ethoxyaniline	0	1
2-Ethoxyaniline	0	1
4-Aminophenol	0	1
3-Aminophenol	0	0
2-Aminophenol	0	0
2-Methoxyaniline	0	1
4-Chloro-1,3-phenylenediamine	1	1
2-Nitro-1,4-phenylenediamine	1	1
4-Nitro-1,3-phenylenediamine	1	1
4-Nitro-1,2-phenylenediamine	1	1

[1] The table reports the mutagenicity of the aromatic and heteroaromatic amines as: 0 = negative; 1 = positive.

[2] TA98 results predicted using topostructural and topochemical indices.

Computation of Indices

Topological indices used in this study have been calculated by POLLY 2.3 [27] which can calculate a total of 102 indices. These indices include Wiener index [28], connectivity indices [11, 12], information theoretic indices defined on distance matrices of graphs [13, 14], a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [15–18], as well as Balaban's *J* indices [19–21]. Table II provides brief definitions for the topological indices included in this study.

TABLE II Symbols, definitions and classifications of topological parameters

	Topostructural
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order $h = 0-6$
${}^h\chi_C$	Cluster connectivity index of order $h = 3-6$
${}^h\chi_{Ch}$	Chain connectivity index of order $h = 3-6$
${}^h\chi_{PC}$	Path-cluster connectivity index of order $h = 4-6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's J index based on distance
	Topochemical
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order $h = 0-6$
${}^h\chi^b_C$	Bond cluster connectivity index of order $h = 3-6$
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order $h = 3-6$
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	Valence path connectivity index of order $h = 0-6$
${}^h\chi^v_C$	Valence cluster connectivity index of order $h = 3-6$
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order $h = 3-6$
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

Values for $\log P$ and the quantum chemical parameters ϵ_{HOMO} and ϵ_{LUMO} were taken from the work of Debnath *et al.* [25]. Octanol/water partition coefficients ($\log P$) were determined experimentally for a set of 67 aromatic and heteroaromatic amines and, when these values were determined to be in agreement with values calculated using the CLOGP program (release

3.54), the remainder of the log *P* values were calculated using CLOGP [29]. The quantum chemical parameters provided by Debnath *et al.*, ϵ_{HOMO} and ϵ_{LUMO} were calculated using the semi-empirical AM1 of MOPAC 4.10 (Quantum Chemistry Program Exchange No. 455) [30].

Data Reduction

Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices and other indices may equal zero.

The set of 95 TIs was partitioned into two distinct sets: 38 topostructural indices and 57 topochemical indices. Topostructural indices are indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. The categorization of the 95 TIs into these sets is shown in Table II.

To further reduce the number of independent variables to be used for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [31]. This variable clustering procedure divides the set of indices into disjoint clusters such that each cluster is essentially unidimensional. The index most correlated with each cluster, as well as any indices which were poorly correlated with the cluster ($r < 0.70$), were selected for model development. Variable clustering was performed independently for both the topostructural and topochemical subsets.

Statistical Analysis and Hierarchical DFA

Selection of indices for the final models was conducted using all subsets regression on the sets of indices chosen through variable cluster analysis in the SAS procedure REG [32]. This all subsets procedure was performed on four distinct sets of indices: (1) the topostructural indices selected by variable clustering, (2) the topostructural indices selected in all subsets regression and the topochemical indices selected during variable clustering, (3) the topostructural and topochemical indices selected in all subsets regression and $\log P$, and 4) the model chosen for topostructural and topochemical indices with $\log P$ and with the addition of ϵ_{HOMO} and ϵ_{LUMO}. These sets of indices were then used to develop and crossvalidate discriminant function models for classifying the mutagenicity/non-mutagenicity of the 127 aromatic and heteroaromatic amines. Figure 1 illustrates the process for the selection of indices and formulation of DFA models.

RESULTS AND DISCUSSION

In the first step of our hierarchical modeling, 38 topostructural parameters were subjected to variable clustering procedure. The following indices were retained from the five clusters generated: $I_D^W, \overline{IC}, O, {}^4\chi_C, {}^6\chi_{Ch}, {}^4\chi_{PC}, P_3, J$. These five clusters explained a total variation of 35.29 and the proportion of the variance explained was equal to 92.86%. Of the 57 topochemical indices, the following ten indices were selected from eight clusters: $IC_0, IC_2, IC_4, SIC_2, SIC_4, {}^4\chi_C^b, {}^6\chi_{Ch}^b, {}^4\chi_{PC}^b, {}^2\chi^v, J^Y$. The eight clusters generated from the topochemical indices resulted in a total variation explained of 51.65 and the proportion of the variance explained was equal to 90.61%. These indices were then included in the all subsets regression procedure for the selection of final indices for discriminant function analysis. In all cases, the RSQUARE and ADJRSQ values were examined as indicators of model fit, however the final models were selected based on the Mallow's *Cp* statistic (CP). Statistics for the cluster analysis and the inter-correlation of the clusters for the topostructural indices are presented in Tables III and IV, respectively. Similar statistics for the variable clustering of the topochemical indices can be found in Tables V and VI.

The all subsets regression of the eight topostructural indices resulted in the selection of the following indices for model development: $I_D^W, \overline{IC}, P_3$. These indices were used to create the topostructural DFA model, the simplest model in the hierarchy, and were also combined with the ten topochemical indices to create the second model in the hierarchy. All subsets regression of the thirteen topostructural and topochemical indices resulted in the selection of the following indices for modeling: $I_D^W, \overline{IC}, P_3, IC_0, SIC_2$. These indices were combined with $\log P$ and resulted in a six parameter model with $\log P$ added to the complete set of descriptors from the second model. Finally, the quantum chemical descriptors, ϵ_{HOMO} and ϵ_{LUMO}, were combined with the set of six indices and all subsets regression was used again

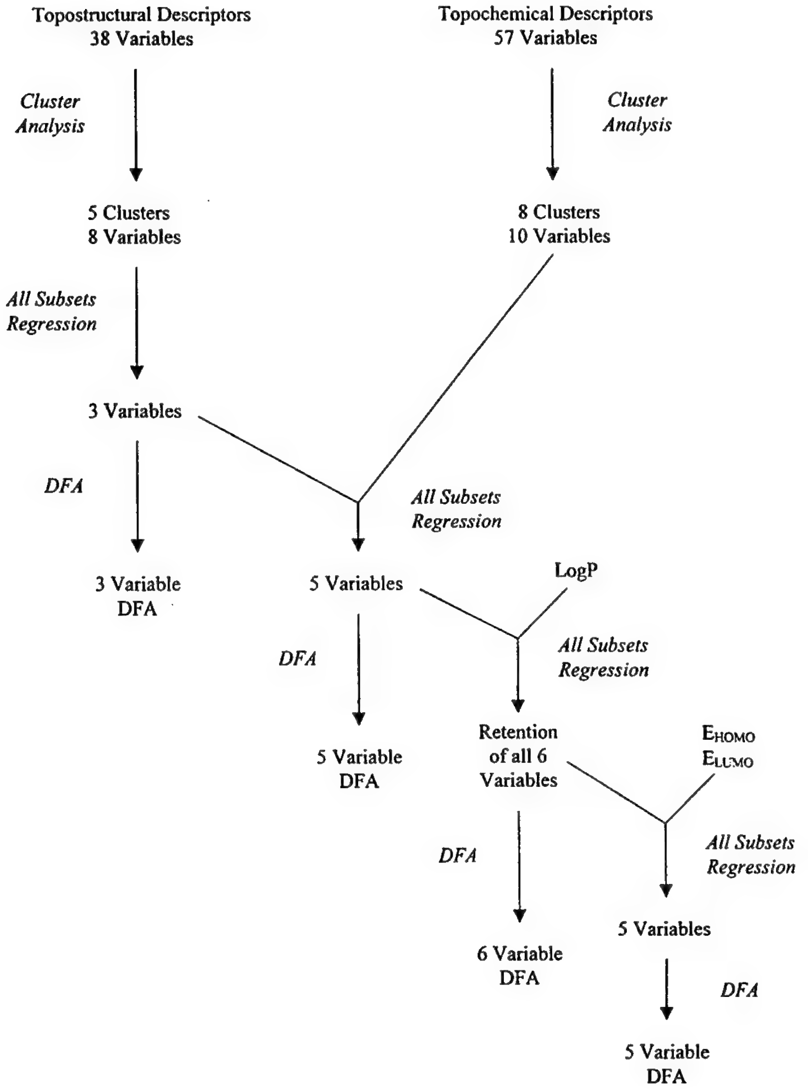


FIGURE 1 Illustration of the hierarchical method of index selection and discriminant function analysis.

to select the best parameters for model construction. This procedure resulted in the selection of the following model: $I_D^W, \overline{IC}, P_3, \log P, \in_{\mathrm{LUMO}}$.

Discriminant function analysis, using the SAS procedure DISCRIM [33], was used to develop models for predicting mutagenicity/non-mutagenicity

TABLE III Statistics for the variable cluster analysis of the topostructural indices

Cluster	*Members*	*Variation explained*	*Proportion explained*	*Second eigenvalue*	*Index most correlated*	*Correlation*
1	18	16.99	0.94	0.71	P_3	0.9918
2	2	2.00	1.00	0.00	${}^4\chi_C$	0.9992
3	3	2.15	0.71	0.72	${}^6\chi_{\mathrm{Ch}}$	0.9104
4	12	11.41	0.95	0.45	I_D^W	0.9977
5	3	2.73	0.91	0.18	${}^4\chi_{\mathrm{PC}}$	0.9474

TABLE IV Intercorrelation of the clusters generated in the variable cluster analysis of the topostructural indices

Cluster	1	2	3	4	5
1	1.0000				
2	0.0735	1.0000			
3	0.6317	−0.0707	1.0000		
4	0.9327	0.1389	0.3922	1.0000	
5	0.7131	0.4006	0.2275	0.7793	1.0000

TABLE V Statistics for the variable cluster analysis of the topochemical indices

Cluster	*Members*	*Variation explained*	*Proportion explained*	*Second eigenvalue*	*Index most correlated*	*Correlation*
1	19	17.61	0.93	0.58	${}^2\chi^v$	0.9686
2	8	7.52	0.94	0.42	SIC_4	0.9876
3	4	3.76	0.94	0.24	${}^4\chi_C^b$	0.9484
4	6	5.11	0.85	0.80	J^Y	0.8889
5	5	4.72	0.94	0.23	IC_4	0.9880
6	4	3.72	0.93	0.27	${}^6\chi_{\mathrm{Ch}}^b$	0.9419
7	6	4.68	0.78	0.79	SIC_2	0.9079
8	5	4.52	0.90	0.21	${}^4\chi_{\mathrm{PC}}^b$	0.9225

TABLE VI Intercorrelation of the clusters generated in the variable cluster analysis of the topochemical indices

Cluster	1	2	3	4	5	6	7	8
1	1.0000							
2	−0.4121	1.0000						
3	0.2311	−0.2150	1.0000					
4	−0.8162	0.4459	−0.0885	1.0000				
5	0.3407	0.6649	−0.0641	−0.2594	1.0000			
6	0.4739	0.2192	−0.0509	−0.4812	0.5033	1.0000		
7	−0.5604	0.4636	−0.1072	0.7565	−0.0130	−0.2089	1.0000	
8	0.7805	−0.5046	0.5542	−0.4287	0.0484	0.1481	−0.2913	1.0000

TABLE VII Results of the cross-validated discriminant function analyses

Hierarchical classes	*Indices*	*% Correct (non-mutagens)*	*% Correct (mutagens)*
Topostructural	$I_D^W, \overline{IC}, P_3$	28.6	95.3
Topostructural + Topochemical	$I_D^W, \overline{IC}, P_3,$ IC_0, SIC_2	42.9	93.4
Topological + log P	$I_D^W, \overline{IC}, P_3,$ $IC_0, SIC_2, \log P$	38.1	95.3
Topological + log P + Quantum chemical	$I_D^W, \overline{IC}, P_3,$ $\log P, \epsilon_{LUMO}$	33.3	95.3

of chemicals in the Ames test. Four distinct models were developed using the indices selected from the all subsets regression procedure as described above. The results in Table VII shows that all four models could predict the mutagenicity of chemicals 93% to 95% of the time whereas they were less effective in predicting non-mutagenicity (29% to 43%).

The addition of topochemical to the set of topostructural indices, resulting in the best predictive model, are shown in Table VII. It is clear from the results that the addition of topochemical indices to the set of topostructural indices did slightly decrease the prediction of mutagenicity. However, there was a significant improvement in the prediction of non-mutagenicity by the addition of topochemical indices to the set of independent variables.

Finally, the addition of log P and quantum chemical indices did not make any improvement in the models. This is in line with our earlier work with physical and biochemical properties which showed that topostructural and topochemical indices explain most of the variance in the data [3, 6–10].

Acknowledgments

This paper is contribution number 232 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force.

References

[1] Hall, L. H. and Story, C. T. (1997). Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological states indices using artificial neural networks. *SAR QSAR Environ. Res.*, **6**, 139–161.

[2] Trinajstić, N., Nikolić, S., Lučić, B., Amić, D. and Mihalić, Z. (1997). The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.*, **37**, 631–638.

[3] Gute, B. D. and Basak, S. C. (1997). Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, **7**, 117–131.

[4] Todeschini, R., Vighi, M., Finizio, A. and Gramatica, P. (1997). 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR QSAR Environ. Res.*, **7**, 173–193.

[5] Guo, M., Xu, L., Hu, C. Y. and Yu, S. M. (1997). Study on structure-activity relationship of organic compounds – Applications of a new highly discriminating topological index. *Math. Chem.* (*MATCH*), **35**, 185–197.

[6] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1998). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, In press.

[7] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **37**, 651–655.

[8] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. In: *Quantitative Structure-Activity Relationships in Environmental Sciences*-7 (Chen, F. and Schüürman, G., Eds.). SETAC Press: Pensacola, FL, pp. 245–261.

[9] Gute, B. D., Grunwald, G. D. and Basak, S. C. (1999). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, In press.

[10] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1996). A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.*, **36**, 1054–1060.

[11] Kier, L. B. and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press: Letchworth, Hertfordshire, U.K.

[12] Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.

[13] Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B. and Basak, S. C. (1984). Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.*, **5**, 581–588.

[14] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **67**, 4517–4533.

[15] Basak, S. C., Roy, A. B. and Ghosh, J. J. (1980). Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In: *Proceedings of the Second International Conference on Mathematical Modelling* (Avula, X. J. R., Bellman, R., Luke, Y. L. and Rigler, A. K., Eds.). University of Missouri-Rolla: Rolla, Missouri, pp. 851–856.

[16] Basak, S. C. and Magnuson, V. R. (1983). Molecular topology and narcosis: A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim. Forsch.*, **33**, 501–503.

[17] Roy, A. B., Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications. In: *Mathematical Modelling in Science and Technology* (Avula, X. J. R., Kalman, R. E., Lipais, A. I. and Rodin, E. Y., Eds.). Pergamon Press: New York, pp. 745–750.

[18] Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med. Sci. Res.*, **15**, 605–609.

[19] Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399–404.

[20] Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure & Appl. Chem.*, **55**, 199–206.

[21] Balaban, A. T. (1986). Chemical graphs. Part 48. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem.* (*MATCH*), **21**, 115–122.

[22] Kier, L. B. and Hall, L. H. (1990). An electrotopological-state index for atoms in molecules. *Pharm. Res.*, **8**, 801–807.

[23] Kier, L. B., Hall, L. H. and Frazer, J. W. (1991). An index of electrotopological state for atoms in molecules. *J. Math. Chem.*, **7**, 229–241.

[24] Balaban, A. T. (1992). Using real numbers as vertex invariants for third-generation topological indices. *J. Chem. Inf. Comput. Sci.*, **32**, 23–28.

[25] Debnath, A. K., Debnath, G., Shusterman, A. J. and Hansch, C. (1992). A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.*, **19**, 37–52.

[26] Benigni, R., Andreoli, C. and Giuliani, A. (1994). QSAR models for both mutagenic potency and activity: Application to nitroarenes and aromatic amines. *Environ. Mol. Mutagen.*, **24**, 208–219.

[27] Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1988). POLLY 2.3: Copyright of the University of Minnesota.

[28] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.

[29] Leo, A. (1988). CLOGP 3.54. Medicinal Chemistry Project, Pomona College, Claremont, CA.

[30] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. and Stewart, J. J. P. (1985). AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, **107**, 3902–3909.

[31] SAS Institute Inc. (1988). The VARCLUS procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 34, pp. 949–965.

[32] SAS Institute Inc. (1988). The REG procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 28, pp. 773–875.

[33] SAS Institute Inc. (1988). The DISCRIM procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 16, pp. 359–447.

APPENDIX 1.3 Hazard assessment modeling: An evolutionary ensemble approach

Hazard Assessment Modeling: An Evolutionary Ensemble Approach

David W. Opitz
Department of Computer Science
University of Montana
Missoula, MT 59812 (USA)
opitz@cs.umt.edu
406-243-2831

Subhash C. Basak **Brian D. Gute**
Natural Resources Research Institute
University of Minnesota
Duluth, MN 55811 (USA)
{sbasak, bgute}@wyle.nrri.umn.edu
218-720-4230

Abstract

This paper presents a novel and effective genetic algorithm approach for generating computational models for hazard assessment. With millions of proposed chemicals being registered each year, it is impossible to come even remotely close to completing the battery of tests needed for the proper understanding of the toxic effects of these chemicals. Computer models can give quick, cheap, and environmentally friendly hazard assessments of chemicals. Our approach works by first extracting a hierarchy of theoretical descriptors of the structure of a compound, then filtering these numerous descriptors with a genetic algorithm approach to ensemble feature selection. We tested the utility of our approach by modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. Our results demonstrate a very important point: that our method is able to accurately predict toxicity directly from structure.

1 INTRODUCTION

By the end of 1998 the number of chemicals registered with the Chemical Abstract Service rose to over 19 million (CAS 1999). This is an increase of over 3 million chemicals between 1996 and 1998. It is desirable to test each of these chemicals for their effects on the environment and human health (which we refer to as *hazard assessment*); however, completing the battery of tests necessary for the proper hazard assessment of even a single compound is a costly and time-consuming process. Therefore, there is simply not enough time or money to complete these test batteries for even a tiny portion of the compounds which are registered today (Menzel 1995). An alternative to these traditional test batteries is to develop computational models for hazard assessment. Computational models are fast (milliseconds per compound), cheap (less than one cent per compound), and do not run the risk of adversely affecting the environment during testing. Additionally, these computational methods can replace or limit the amount of animal testing that is necessary. Thus computational models can easily process *all* registered chemicals and flag the ones that require further testing. The central problem with this approach is developing class specific models that can be considered accurate enough to be useful. In this paper, we present a novel and effective approach for learning computational hazard assessment models by using an ensemble feature selection algorithm based on genetic algorithms (GAs) to filter numerous theoretical descriptors of chemical structure.

To better illustrate the need for effective and quick hazard assessment, we should consider the situation of the industrial chemicals "grandfathered" into continued use under the Toxic Substances Control Act (TSCA) of 1976. TSCA has required that a suite of physicochemical and toxicological screens be run on all commercial compounds (those produced or imported in volumes exceeding one million pounds annually) developed after 1976. However, there are almost 3,000 chemicals that were "grandfathered" in with the understanding that it would be the responsibility of the chemical manufacturing industry to ultimately supply information about these chemicals. Only recently, after a 20-year delay, are the chemical manufacturers talking about running 2,800 of these compounds through basic toxicity screens and while this is promising, these screens will not be completed until 2004 and at a cost of between $500 to $700 million dollars. So it will be another five years before we have basic toxicity data on compounds that have been in wide-spread use for more than twenty years (Johnson 1998).

One of the fundamental principles of biochemistry is that activity is dictated by structure (Hansch 1976). Following this principle, one can use theoretical molecular descriptors that quantify structural aspects of a molecule to quantitatively determine its activity (Basak & Grunwald 1995; Cramer, Famini, & Lowrey 1993). These theoretical descriptors can be generated directly from the known structure of the molecule and used to estimate its properties, without the need for further experimental data. This is important due to that fact that, with chemicals needing to be evaluated for hazard assessment, there is a scarcity of available experimental data that is normally required as inputs (i.e., independent variables) to traditional quantitative structure-activity relationship (QSAR) model development. A QSAR model based solely on theoretical descriptors on the other hand can process all registered chemicals for hazard assessment.

Our hierarchical approach examines the relative contributions of theoretical descriptors of gradually increasing complexity (structural, chemical, shape, and quantum chemical descriptors). This approach is important as none of the individual classes of parameters are very effective at predicting toxicity (Gute & Basak 1997); however, we show in this paper that we can effectively predict toxicity if we combine all levels of descriptors. One potential problem with using our hierarchical approach is that it often gives many independent variables as compared to data points since having a limited number of data points in not uncommon in hazard assessment. For instance, in our case study of predicting acute toxicity (LC_{50}) of benzene derivatives, we have 95 independent variables and 69 data points. Therefore, reducing the number of independent variables is critical when attempting to model small data sets. The smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors). In some of our earlier QSAR studies we have used statistical methods such as principal components analysis (PCA) and variable clustering methods to reduce the number of independent variables (Basak & Grunwald 1995; Gute & Basak 1997; Gute, Grunwald, & Basak In press).

As an alternative solution, we use our previous ensemble feature selection approach (Opitz 1999) that is based on GAs. An "ensemble" is a combination of the outputs from a *set* of models that are generated from separately trained inductive learning algorithms. Ensembles have been shown to, in most cases, greatly improve generalization accuracy over a single learning model (Breiman 1996; Maclin & Opitz 1997; Shapire *et al.* 1997). Recent research has shown that an effective ensemble should consist of a set of models that are not only highly correct, but ones that make their errors on different parts of the input space as well (Hansen & Salamon 1990; Krogh & Vedelsby 1995; Opitz & Shavlik 1996a). Varying the feature subsets used by each member of the ensemble helps promote the necessary diversity and create a more effective ensemble (Opitz 1999). We use GAs to search through the enormous space of finding a set of feature subsets that will promote disagreement among the component members of an ensemble while still maintaining the component member's accuracy.

Combining our approach of generating hierarchical theoretical descriptors with our other approach to GA-based ensemble feature selection, we are able to generate an effective model for predicting the toxicity of benzene derivatives using only a few compounds. Our results show that our model is nearly as accurate as the battery of tests necessary for the proper hazard assessment of a single compound. Our results also confirm that our new ensemble feature selection approach is more effective than previous approaches for modeling hazard assessment.

The rest of the paper is organized as follows. First we provide background and related work for both our hierarchical QSAR approach and our GA-based ensemble feature selection approach. This is followed by results of our approach applied to benzene derivatives. Finally, we discuss these results and provide future work.

2 QSAR AND THEORETICAL METHODS

QSARs have come into widespread use for the prediction of various molecular properties, as well as biological, pharmacological and toxicological responses. Traditional QSAR techniques use empirical properties (Dearden 1990; Hansch & Leo 1995; de Waterbeemd 1995); however, due to the scarcity of available data for the majority of chemicals needing to be evaluated for hazard assessment, these physicochemical properties necessary for traditional QSAR model development may not be available. When this is the case, it is imperative that there are methods available which make use of nonempirical parameters, which we term theoretical molecular descriptors.

Topological indices (TIs) are numerical graph invariants that quantify certain aspects of molecular structure (Gute & Basak 1997; Gute, Grunwald, & Basak In press). The different classes of TIs provide us with nonempirical, quantitative descriptors that can be used in place of experimentally derived descriptors in QSARs for the prediction of properties.

Our recent studies have focused on the role of different classes of theoretical descriptors of increasing levels of complexity and their utility in QSAR (Gute & Basak 1997; Gute, Grunwald, & Basak In press). Four distinct sets of theoretical descriptors have been used in this study: topostructural, topochemical, geometric, and quantum chemical indices. Gute and Basak 1997 provide the detailed list of the indices included in our study.

2.1 TOPOLOGICAL INDICES

The topostructural and topochemical indices fall into the category normally considered topological indices. Topostructural indices (TSIs) are topological indices that only encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs), irrespective of the chemical nature of the atoms involved in bonding or factors such as hybridization states and the number of core/valence electrons in individual atoms. Topochemical indices (TCIs) are parameters that quantify information regarding the topology (connectivity of atoms), as well as specific chemical properties of the atoms comprising a molecule. These indices are derived from weighted molecular graphs where each vertex (atom) or edge (bond) is properly weighted with selected chemical or physical property information.

The complete set of topological indices used in this study, both the topostructural and the topochemical, have been calculated using POLLY 2.3 (Basak, Harriss, & Magnuson 1988) and software developed by the authors. These indices include the Wiener index (Wiener 1947), the connectivity indices developed by Randic 1975 and higher order connectivity indices formulated by Kier and Hall 1986, bonding connectivity indices defined by Basak and Magnuson 1988, a set of information theoretic indices defined on the distance matrices of simple molecular graphs (Hansch & Leo 1995), and neighborhood complexity indices of hydrogen-filled molecular graphs, and Balaban's 1983 J indices.

2.2 GEOMETRICAL INDICES

The geometrical indices are three-dimensional Wiener numbers for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume. Van der Waals volume, V_W (Bondi 1964), was calculated using Sybyl 6.1 from Tripos Associates, Inc. of St. Louis. The 3-D Wiener numbers were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed in our lab (SYBYL 1998). Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using CONCORD 3.0.1 from Tripos Associates, Inc. Two variants of the 3-D Wiener number were calculated: ${}^{3D}W_H$ and ${}^{3D}W$. For ${}^{3D}W_H$, hydrogen atoms are included in the computations and for ${}^{3D}W$ hydrogen atoms are excluded from the computations.

2.3 QUANTAM CHEMICAL PARAMETERS

The following quantum chemical parameters were calculated using the Austin Model version one (AM1) semi-empirical Hamiltonian: energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital (E_{HOMO1}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second lowest unoccupied molecular orbital (E_{LUMO1}), heat of formation (ΔH_f), and dipole moment (μ). These parameters were calculated using MOPAC 6.00 in the SYBYL interface (Stewart 1990).

3 FILTERING DESCRIPTORS

As stated above, one potential problem with including all theoretical descriptors in the hierarchy is that it gives many independent variables when compared to the limited number of data points available for hazard assessment modeling of a particular chemical derivative. Compounding this problem is that a salient descriptor for one hazard assessment model may not be a salient descriptor for another problem. That is, the relevance of a descriptor for predicting hazard assessment is often problem dependent. This section describes our approach for automatically filtering the descriptors with a GA-based approach to ensemble feature detection. Before explaining our algorithm, we briefly cover the notion of ensembles.

3.1 ENSEMBLES

Figure 1 illustrates the basic framework of a predictor ensemble. Each predictor in the ensemble (predictor 1 through predictor N in this case) is first trained using the training instances. Then, for each example, the predicted output of each of these predictors (o_i in Figure 1) is combined to produce the output of the ensemble ($\hat{o}$ in Figure 1). Many researchers (Breiman 1996; Hansen & Salamon 1990; Krogh & Vedelsby 1995;

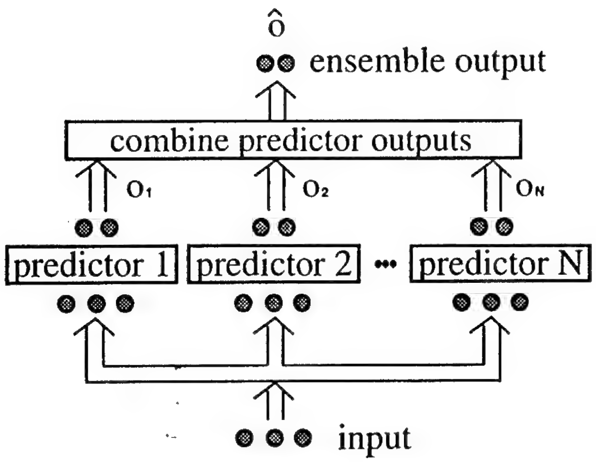


Figure 1: A predictor ensemble.

Opitz & Shavlik 1997) have demonstrated the effectiveness of combining schemes that are simply the weighted average of the predictors (i.e., $\hat{o} = \sum_{i \in N} w_i \cdot o_i$ and $\sum_{i \in N} w_i = 1$), and this is the type of ensemble on which we focus in this article.

Combining the output of several predictors is useful only if there is disagreement on some inputs. Obviously, combining several identical predictors produces no gain. Hansen and Salamon 1990 proved that for an ensemble, if the average error rate for an example is less than 50% and the predictors in the ensemble are independent in the production of their errors, the expected error for that example can be reduced to zero as the number of predictors combined goes to infinity; however, such assumptions rarely hold in practice.

Krogh and Vedelsby 1995 later proved that the ensemble error can be divided into a term measuring the average generalization error of each individual predictor and a term called diversity that measures the disagreement among the predictors. Formally, they define the diversity term, d_i, of predictor i on input x to be:

$$d_i(x) \equiv [o_i(x) - \hat{o}(x)]^2. \tag{1}$$

The quadratic error of predictor i and of the ensemble are, respectively:

$$\epsilon_i(x) \equiv [o_i(x) - f(x)]^2, \tag{2}$$

$$e(x) \equiv [\hat{o}(x) - f(x)]^2, \tag{3}$$

where $f(x)$ is the target value for input x. If we define $\hat{E}$, E_i, and D_i to be the averages, over the input distribution, of $e(x)$, $\epsilon(x)$, and $d(x)$ respectively, then the ensemble's generalization error can be shown to consist of two distinct portions:

$$\hat{E} = \bar{E} - \bar{D}, \tag{4}$$

where $\bar{E}$ ($= \sum_i w_i E_i$) is the weighted average of the individual predictor's generalization error and $\bar{D}$ ($= \sum_i w_i D_i$) is the weighted average of the diversity among these predictors. What the equation shows then, is that an ideal ensemble consists of highly correct predictors that disagree as much as possible. Opitz and Shavlik 1996a; 1996b empirically verified that such ensembles generalize well.

Regardless of theoretical justifications, methods for creating ensembles center around producing predictors that disagree on their predictions. Generally, these methods focus on altering the training process in the hope that the resulting predictors will produce different predictions. For example, neural network techniques that have been employed include methods for training with different topologies, different initial weights, different parameters, and training only on a portion of the training set (Alpaydin 1993; Freund & Schapire 1996; Hansen & Salamon 1990; Maclin & Shavlik 1995).

Numerous techniques try to generate disagreement among the classifiers by altering the training set each classifier sees. The two most popular techniques are Bagging (Breiman 1996) and Boosting (Freund & Schapire 1996). Bagging is a bootstrap ensemble method that trains each network in the ensemble with a different partition of the training set. It generates each partition by randomly drawing, with replacement, N examples from the training set, where N is the size of the training set. As with Bagging, Boosting also chooses a training set of size N and initially sets the probability of picking each example to be $1/N$. After the first network, however, these probabilities change to emphasize misclassified instances. A large number of extensive empirical studies have shown that these are highly successful methods that nearly always generalize better than their individual component predictors (Bauer & Kohavi 1998; Maclin & Opitz 1997; Quinlan 1996). Neither approach is appropriate for our domain since we are data poor and cannot afford to waste training examples; however, we are feature rich and can afford to create diversity by instead varying the inputs to the learning algorithms. *Varying the feature subsets to create a diverse set of accurate predictors is the focus of the next section.*

3.2 THE GEFS ALGORITHM

The goal of our algorithm is to find a set of feature subsets that creates an ensemble of classifiers (neural networks in this study) that maximize equation 1 while minimizing equation 2. The space of candidate sets is enormous and thus is particularly well suited for ge-

Table 1: The GEFS algorithm.

GOAL: Find a set of input subsets to create an accurate and diverse classifier ensemble.

1. Using varying inputs, create and train the initial population of classifiers.
2. Until a stopping criterion is reached:
   (a) Use genetic operators to create new networks.
   (b) Measure the diversity of each network with respect to the current population.
   (c) Normalize the accuracy scores and the diversity scores of the individual networks.
   (d) Calculate fitness of each population member.
   (e) Prune the population to the N fittest networks.
   (f) Adjust λ.
   (g) The current population is the ensemble.

netic algorithms. Table 1 summarizes our recent algorithm (Opitz 1999) called GEFS (for Genetic Ensemble Feature Selection) that uses GAs to generate a set of classifiers that are accurate and diverse in their predictions. GEFS starts by creating and training its initial population of networks. The representation of each individual of our population is simply a dynamic length string of integers, where each integer indexes a particular feature. We create networks from these strings by first having the input nodes match the string of integers, then creating a standard single-hidden-layer, fully connected neural network. Our algorithm then creates new networks by using the genetic operators of crossover and mutation.

GEFS trains these new individuals using backpropogation. It adds new networks to the population and then scores each population member with respect to its prediction accuracy and diversity. GEFS normalizes these scores, then defines the fitness of each population member (i) to be:

$$Fitness_i = Accuracy_i + \lambda\ Diversity_i \qquad (5)$$

where λ defines the tradeoff between accuracy and diversity. Finally, GEFS prunes the population to the N most-fit members, then repeats this process. At every point in time, the current ensemble consists of simply averaging (with equal weight) the predictions of the output of each member of the current population. Thus as the population evolves, so does the ensemble.

We define accuracy to be network i's training-set accuracy. (One may use a validation-set if there are enough training instances.) We define diversity to be the average difference between the prediction of our component classifier and the ensemble. We then separately normalize both terms so that the values range from 0 to 1. Normalizing both terms allows λ to have the same meaning across domains.

It is not always clear at what value one should set λ; therefore, we automatically adjust λ based on the discrete derivatives of the ensemble error $\hat{E}$, the average population error $\bar{E}$, and the average diversity $\bar{D}$ within the ensemble. First, we never change λ if $\hat{E}$ is decreasing; otherwise we (a) increase λ if $\bar{E}$ is not increasing and the population diversity $\bar{D}$ is decreasing; or (b) decrease λ if $\bar{E}$ is increasing and $\bar{D}$ is not decreasing. We started λ at 1.0 for the experiments in this article. The amount λ changes is 10% of its current value.

We create the initial population by randomly choosing the number of features to include in each feature subset. For classifier i, the size of each feature subset (N_i) is independently chosen from a uniform distribution between 1 and twice the number of original features in the dataset. We then randomly pick, with replacement, N_i features to include in classifier i's training set. Note that some features may be picked multiple times while others may not be picked at all; replicating inputs for a neural network may give the network a better chance to utilize that feature during training. Also, replicating a feature in a genome encoding allows that feature to better survive to future generations.

Our crossover operator uses dynamic-length, uniform crossover. In this case, we chose the feature subsets of two individuals in the current population proportional to fitness. Each feature in both parent's subset is independently considered and randomly placed in the feature set of one of the two children. Thus it is possible to have a feature set that is larger (or smaller) than the largest (or smallest) of either parent's feature subset. Our mutation operator works much like traditional genetic algorithms; we randomly replace a small percentage of a parent's feature subset with new features. With both operators, the network is trained from scratch using the new feature subset; thus no internal structure of the parents are saved during the crossover.

4 RESULTS

We tested the utility of combining our approach for generating numerous hierarchical theoretical descriptors of compounds with our approach for filtering these descriptors with GEFS by modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. The data was taken from the work of Hall, Kier and Phipps 1984 where acute aquatic toxicity was measured in fathead minnow (*Pimephales promelas*). Their data was compiled from eight other sources, as well as some original work which was conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota. This set of chemicals was composed of benzene and 68 substituted benzene derivatives.

Table 2 gives our results. We studied three approaches for modeling toxicity: (1) giving all theoretical descriptors to a neural network, (2) reducing the feature set in a traditional previously published (Gute & Basak 1997) manner, and (3) using our new genetic algorithm technique on the entire feature set to create a neural network ensemble. Results for our approaches are from leave-one-out experiments (i.e., 69 training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (68 in this case) to the learning algorithms for training. (It is worth noting that each member of the ensemble sees the same 68 training instances for each training/test set partition and thus ensembles have no unfair advantage over other learners.) This process is repeated 69 times so that each example is a part of the test set once and only once. Leave-one-out tests *generalization* accuracy of a learner, whereas training set accuracy tests only the learner's ability to memorize. Generalization error from the test set is the true test of accuracy and is what we report here.

We first trained neural networks using all 95 parameters. The networks contained 15 hidden units and we trained the networks for 1000 epochs. We normalized each input parameter to a values between 0 and 1 before training. Additional parameter settings for the neural networks included a learning rate of 0.05, a momentum term of 0.1, and weights initialized randomly between -0.25 and 0.25. With all 95 input parameters, the neural networks obtained a test-set correlation coefficient between predicted toxicity and measured toxicity (explained variance) of $R^2 = 0.868$ and a standard error of 0.29. Target toxicity measurements ranged from 3.04 to 6.37.

Our first method for feature-set reduction follows the work of Gute and Basak 1997 on toxicity domains. Their method begins by using the VARCLUS method of SAS 1998 to select subsets of topostructural and topochemical parameters for QSAR model development. With this method, the set of topological indices is first partitioned into two distinct sets, the topostructural indices and the topochemical indices.

Table 2: Relative effectiveness of statistical and neural network methods in estimating LC_{50} of 69 benzene derivatives.

Method	R^2	Standard Error
NN with 95 inputs	0.868	0.29
VARCLUS	0.825	0.32
NN with GEFS	0.893	0.27

To further reduce the number of independent variables for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the VARCLUS procedure. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with their cluster ($R^2 < 0.70$). The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure resulted in a set of five topostructural indices and a set of nine topochemical indices. These indices were combined with the three geometric and six quantum chemical parameters described earlier. Their approach then applied linear regression to these 23 parameters. This study found that an accurate linear regression model for acute aquatic toxicity required descriptors from all four levels of the hierarchy: topostructural, topochemical, geometrical and quantum chemical. This model utilized seven descriptors and obtained an explained variance (R^2) of 0.863 and a standard error of 0.30 on the whole data set used as a training set. Our leave-one-out experiment gave an $R^2 = 0.825$ and a standard error of 0.32.

Finally we applied our genetic algorithm technique, GEFS, using all 95 parameters. The parameter settings for the networks in the ensemble were the same as the settings for the single networks in the first experiment. Parameter settings for the genetic algorithm portion of GEFS includes a mutation rate of 50%, a population size of 20, a $\lambda = 1.0$, and a search length of 100 networks (20 networks for the initial population and 80 networks created from crossover and mutation). While the mutation rate may seem high as compared with traditional genetic algorithms, certain aspects of our approach call for a higher mutation rate (such as the criterion of generating a population that cooperates as well as our emphasis on diversity); other mutation values were tried during our pilot studies. With this approach, we obtained a test-set correlation coefficient of $R^2 = 0.893$ and a standard error of 0.27; the initial population of 20 networks obtained a test-set

$R^2 = 0.835$ and a standard error of 0.31.

5 DISCUSSION AND FUTURE WORK

The correlation coefficient between the predicted value from the computational model and the target value derived from the toxicity test is an extremely informative metric of accuracy in this case. The exact numeric value of most toxicity tests is not as important as the relative ordering and spread of these values. Thus, a perfect correlation ($R^2 = 1.0$) between the computation model and target toxicity shows the computational model is as informative as the toxicity obtained from a battery of expensive and time-consuming tests – regardless of the standard error. Note the standard error of 0.27 is fairly good, given the toxicity measurements ranged from 3.04 to 6.37.

While the neural network technique and the standard data-reduction technique obtained decent correlation with measured toxicity, our ensemble technique was about 20% closer to perfect correlation. Note that GEFS produces an accurate initial population and that running GEFS longer with our genetic operators can further increase performance. Thus our approach can be viewed as an "anytime" learning algorithm. Such a learning algorithm should produce a good concept quickly, then continue to search concept space, reporting the new "best" concept whenever one is found (Opitz & Shavlik 1997). This is important since, for most hazard assessment, an expert is willing to wait for days, or even weeks, if a learning system can produce an improved model for predicting toxicity.

Our results demonstrate a very important point: that our method is able to accurately predict toxicity directly from structure. Compared to the actual battery of tests necessary to measure toxicity, a computer model is much cheaper, much faster, and does not have a negative impact on the environment. It is important to also note that the computer model does not have to be the final measurement for hazard assessment; additional tests can be run on compounds that are either flagged by the model, or require more tests by the nature of their use (such as a benzene derivative that may become a standard fuel). Not only can good computer models become filters, they will probably be the only viable option for processing all registered chemicals.

While the method proposed here has proven effective, there is much future work that needs to be completed. For instance, we plan to test our method on other data sets of chemical derivatives; investigate other ensemble feature selection techniques; investigate variants to our genetic algorithm approach, and finally investigate the utility of other descriptors, such as bio-descriptors.

6 CONCLUSIONS

In this paper we presented a novel approach for creating a computer model for hazard assessment. Our approach works by first extracting a hierarchy of theoretical descriptors derived from the structure of a compound, then filtering the numerous possible descriptors with a genetic algorithm approach to ensemble feature selection. We tested the utility of our approach by modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. Our results demonstrate the ability of our approach to accurately predict toxicity directly from structure. Thus our new algorithm further increases the applicability of computer models to the problem of predicting chemical activity directly from its structure.

Acknowledgements

This work was partially supported by National Science Foundation grant IRI-9734419, a University of Montana MONTS grant, U.S. Air Force grant F49620-96-1-0330, and is contribution number 246 for the Center for Water and the Environment of the Natural Resources Research Institute.

References

Alpaydin, E. 1993. Multiple networks for function learning. In *Proceedings of the 1993 IEEE International Conference on Neural Networks*, volume I, 27–32. San Fransisco: IEEE Press.

Balaban, A. 1983. Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* 55:199–206.

Basak, S., and Grunwald, G. 1995. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry* 19:231–237.

Basak, S., and Magnuson, V. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* 19:17–44.

Basak, S.; Harriss, D.; and Magnuson, V. 1988. Polly 2.3. Copyright of the University of Minnesota.

Bauer, E., and Kohavi, R. 1998. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning.*

Bondi, A. 1964. Van der waals volumes and radii. *J. Phys. Chem.* 68:441–451.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

CAS. 1999. The latest cas registry number and substance count. http://www.cas.org/cgi-bin/regreport.pl.

Cramer, C.; Famini, G.; and Lowrey, A. 1993. Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Acc. Chemical Research* 26:599–605.

de Waterbeemd, H. V. 1995. Discriminant analysis for activity prediction. In *Chemometric Methods in Molecular Design*, 283–294. VCH Publishers, Inc.

Dearden, J. 1990. Physico-chemical descriptors. In *Environmental Chemistry and Toxicology*, 25–59. Kluwer Academic Publisher.

Freund, Y., and Schapire, R. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156. Morgan Kaufmann.

Gute, B., and Basak, S. 1997. Predicting acute toxicity (LC50) of benzen derivatives using theoretical molecular descripors: A hierarchical QSAR approach. *SAR and QSAR in Environmental Research* 7:117–131.

Gute, B.; Grunwald, G.; and Basak, S. In press. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. In *SAR and QSAR in Environmental Research.*

Hall, L.; Kier, L.; and Phipps, G. 1984. Structure-activity relationship studies on the toxicities of benzene derivatives: I. an additivity model. *Environ. Toxicol. Chem.* 3:355–365.

Hansch, C., and Leo, A. 1995. Exploring QSAR: Fundamentals and applications in chemistry and biology. *American Chemical Society* 557.

Hansch, C. 1976. On the structure of medicinal chemistry. *Journal of Medicinal Chemistry* 19:1–6.

Hansen, L., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.

Johnson, J. 1998. Pact triggers tests: Thousands of chemicals may be tested under toxicity screening program. *Chemical Engineering News* 76(44):19–20.

Kier, L., and Hall, L. 1986. *Molecular Connectivity in Structure-Activity Analysis.* Hertfordshire, UK: Research Studies Press.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. In Tesauro, G.; Touretzky, D.; and Leen, T., eds., *Advances in Neural Information Processing Systems*, volume 7, 231–238. Cambridge, MA: MIT Press.

Maclin, R., and Opitz, D. 1997. An empirical evaluation of bagging and boosting. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 546–551. Providence, RI: AAAI/MIT Press.

Maclin, R., and Shavlik, J. 1995. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.*

Menzel, D. 1995. Extrapolating the future: research trends in modeling. *Toxicology Letters* 79:299–303.

Opitz, D., and Shavlik, J. 1996a. Actively searching for an effective neural-network ensemble. *Connection Science* 8(3/4):337–353.

Opitz, D., and Shavlik, J. 1996b. Generating accurate and diverse members of a neural-network ensemble. In Touretsky, D.; Mozer, M.; and Hasselmo, M., eds., *Advances in Neural Information Processing Systems*, volume 8. Cambridge, MA: MIT Press.

Opitz, D., and Shavlik, J. 1997. Connectionist theory refinement: Searching for good network topologies. *Journal of Artificial Intelligence Research* 6:177–209.

Opitz, D. 1999. Feature selection for ensembles. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence.*

Quinlan, J. R. 1996. Bagging, boosting, and c4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 725–730. AAAI/MIT Press.

Randic, M. 1975. On characterization of molecular branching. *Journal of American Chemical Society* 97:6609–6615.

SAS. 1998. Cary, NC: SAS Institute Inc. chapter SAS/STAT User's Guide, Release 6.03 Edition.

Shapire, R.; Freund, Y.; Bartlett, P.; and Lee, W. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 322–330. Nashville, TN: Morgan Kaufmann.

Stewart, J. 1990. Mopac version 6.00. qcpe #455. US Air Force Academy, CO: Frank J. Seiler Research Laboratory.

SYBYL. 1998. Sybyl version 6.1. Tripos Associates, Inc.

Wiener, H. 1947. Structural determination of paraffin boiling points. *Journal of Am. Chem. Soc.* 69:17–20.

APPENDIX 1.4 Information theoretic indices of neighborhood complexity and their applications

[36] Hall, L.H. and Story, C.T. (1995). Boiling point and critical temperature of a heterogeneous data set: QSAR with atom-type E-State indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **36,** 1004–1014.

[37] Fulcrand, P., Berge, G., Noel, A.-M., Chevallet, P., Castel, J. and Orzalesi, H. (1978). Hydrazide inhibitors of monoamine oxidase: Correlations with Hansch and Free-Wilson. *Eur. J. Med. Chem.* **13,** 177–182.

[38] Tamm, I., Folkers, K., Shunk, C., and Horofall, F.L. (1953). Inhibition of influenza virus multiplication by alkyl derivatives of benzimidazoles. *J. Exp. Med.* **98,** 219–229.

[39] Gough, J. and Hall, L.H. Modeling the toxicity of amide herbicides using the electrotopological state. *Environ. Toxical. Chem.* (in press).

[40] Zakarya, D., Larfaoui, E.M., Boulaamail, A., and Lakhlifi, T. (1996). Analysis of structure relationships for a series of amide herbicides using statistical methods and neural networks. *SAR QSAR Environ. Res.* **5,** 269–279.

[41] Walters, W.P., Stahl, M.T., and Murcko, M. (1998). Virtual screening – an overview. *Drug Discovery Topics* **3,** 260–278.

[42] Van Drie, J. and Lajiness, M. (1998). Approaches to virtual library design. *Drug Discovery Topics* **3,** 274–283.

12. INFORMATION THEORETIC INDICES OF NEIGHBORHOOD COMPLEXITY AND THEIR APPLICATIONS

S.C. Basak

Center for Water and the Environment,
Natural Resources Research Institute,
University of Minnesota, Duluth, Minnesota, USA

A contemporary interest in chemical graph theory and mathematical chemistry is the characterization of molecular structure using graph invariants. Numerical graph invariants or topological indices are derived from molecular graphs and quantify various aspects of molecular architecture including shape, size, complexity, branching, etc.

Information-theoretic or complexity indices have been developed by various authors. These indices are derived from different graph theoretic representations of molecular structure, viz., planar graphs, multigraphs, etc. Our research team has been involved in the formulation of novel information-theoretic parameters, viz., information content (IC), structural information content (SIC), relative nonstructural information content (RNSIC) and complementary information content (CIC), which quantify the degree of heterogeneity and redundancy of topological neighborhoods of atoms in molecules. Such indices can be calculated for different orders of neighborhoods of atoms.

We have applied this novel class of indices in the discrimination of closely related structures, quantification of intermolecular similarity, and

In, *Topological Indices and Related Descriptors in QSAR and QSPR* (J. Devillers and A.T. Balaban, Eds.) Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 563–593

in quantitative structure–property/activity relationship (QSPR/QSAR) studies. This chapter will present the mathematical basis and methods of computation of such indices. We will also undertake a critical analysis of the utility of such indices in characterization of structure and QSAR/QSPR studies.

INTRODUCTION

A recent trend in mathematical chemistry, chemical graph theory, quantitative structure–activity relationship (QSAR) studies as well as predictive toxicology is the use of graph theoretical invariants for the characterization of structure and prediction of properties. Graph theoretic indices have been used for isomer discrimination and characterization of structures [1–3], ordering of sets of closely related molecules as well as prediction of physicochemical, biomedicinal and toxicological properties [4–14]. In environmental toxicology, for example, the Toxic Substances Control Act (TSCA) Inventory has more than seventy six thousand entries. Most of these chemicals do not have the experimental data necessary for their hazard assessment [15–17]. The Chemical Abstracts Service (CAS) database currently has more than sixteen million entries. Most of these chemicals have very little experimental data necessary for the prediction of their therapeutic and toxic potential. In modern combinatorial chemistry, one can produce very large real or virtual libraries very fast. But most of those chemicals do not have any experimental data that is needed to predict their therapeutic or toxic properties. The molecular structure is the only property available for such chemicals. Therefore, nonempirical parameters of chemical structure derived from graph theoretic formalism are being used more frequently by many researchers in QSAR studies pertaining to molecular design, pharmaceutical drug design, and environmental hazard assessment of chemicals [4–14].

A molecular graph represents the topology of the chemical species. Various matrices, e.g., the adjacency matrix, the distance matrix, can be used to symbolize such graphs. During the last two decades various authors have developed numerous graph invariants from such matrices. A graph invariant is a graph theoretic property or parameter which has the same value for isomorphic graphs [18]. A graph invariant is called a topological index (TI) when it is a single number.

Randić's connectivity index [3], the higher order indices defined by Kier *et al.* [19], and the *J* index of Balaban [2] are derived from adjacency and distance matrices of molecular graphs. On the other hand, information theoretic formalism has been applied on matrices as well as neighborhood of vertices of chemical graphs to derive a new class of TIs which are useful in the characterization of structure and prediction of properties. Our research group, in particular, has pioneered the development and applications of novel classes of indices called information theoretic indices of molecular graphs [10, 20, 21]. This chapter will review the philosophical basis, the mathematical background and the utility of these parameters called neighborhood complexity indices.

NEIGHBORHOOD COMPLEXITY INDICES

Graph Theoretic Background

A graph $G = [V, E]$ consists of a finite nonempty set V of points together with a prescribed set E of unordered pairs of distinct points of V. A *structural model* assigns to the points of G a realization in some applied field and each element of E indicates a pair of points which are in the finite nonempty irreflexive symmetric binary relation described by G. In a molecular graph (the conventional chemical structure), the set of atoms comprises the point set V and covalent chemical bonds are elements of E. Such a graph retains the full topology of the molecule and represents molecular structure, where the word "structure" is used to denote a formal system of relations of certain logical types without emphasizing the entities to which they relate. It is because of this general nature that graph-theoretic methods have been used for characterizing structure in such diverse areas as theoretical physics, chemistry, biological, and social sciences, engineering, computer science, and linguistics.

In chemistry, two types of graphs, *viz.*, hydrogen-suppressed graphs and hydrogen-filled graphs, are often used to model molecular structure. While in the former only the non-hydrogen atoms are represented by points, in the latter all atoms (including hydrogen atoms) are represented by vertices. G_1 and G_2 are the hydrogen-suppressed graph and hydrogen-filled graph, respectively, of 2-methyl propane (isobutane in Figure 1).

Such a graph represents the "topology of a molecule" in the sense that it depicts the pattern of connectedness of atoms in the molecule, being, at the same time, independent of such metric aspects of molecular structure as equilibrium distance between nuclei, bond angles, etc.

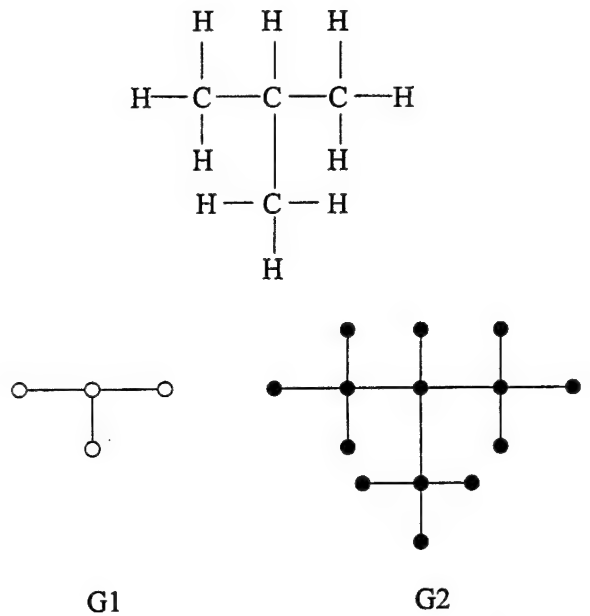


Figure 1 Hydrogen-suppressed and hydrogen-filled graphs of isobutane.

For a long time, chemists have primarily relied upon visual perception in order to relate various aspects of constitutional graphs (structure) to observable chemical phenomena. But a clear and quantitative understanding of the structural basis of properties of molecules necessitates the use of precise mathematical techniques. For the purpose of studying the relationship between chemical structure and property, the bonding topology of a molecule is converted into an expression which may be a matrix, a polynomial, a sequence of numbers or a numerical index. Such a numerical index characterizing chemical structure is called a topological index.

Molecular complexity indices are calculated by the application of information—theoretic formalism on chemical graphs. The science of information theory has grown mainly out of the pioneering studies of Shannon [22], Wiener [23], Ashby [24], and Kolmogorov [25]. There is more than one version of information theory. In Shannon's statistical information theory, information is measured as reduced uncertainty of the system. In the algorithmic theory of Kolmogorov, the quantity of information is defined as the minimal length of a program which allows a one-to-one transformation of an object (set) into another. In applying information theory to chemical structure, we look upon information content of a molecule as a measure of variety or heterogeneity of the system as suggested by Ashby [24].

Molecular complexity indices may be broadly divided into two groups: (1) topological complexity indices, and (2) chemical complexity indices. We believe that such a classification brings the large number of complexity parameters into a proper perspective and helps to rationalize their nature. *Topological complexity indices* are calculated from linear graphs or multigraphs of chemical species where all atoms are represented by points which are indistinguishable from one another. This group includes parameters developed by Rashevsky [26], Bonchev and Trinajstić [27] and Bertz [28]. Successful applications of this class of indices in QSAR studies, discrimination of isomers and chemical synthesis design indicate that they quantitate important aspects of chemical structure.

In many cases, however, it is the nature of an atom or a group of atoms that makes a significant difference in physiochemical/biological properties of molecules. For example, methylene chloride, bromochloromethane and methylene bromide have identical hydrogen-suppressed and hydrogen-filled graphs. On the other hand, while methylene chloride is the least mutagenic of the three dihalomethane compounds, methylene bromide has the highest mutagenic potency. This trend in mutagenicity reflects changes in chemical reactivity arising out of the presence of different heteroatoms. Topological complexity indices are insensitive to effects of this type.

By the term *chemical complexity index* we mean a complexity parameter which takes into account the chemical nature of individual atoms and the bonding patterns in the molecule. This can be accomplished either in terms of bonding topology of weighted graphs or through the use of physicochemical/geometrical characteristics of various atoms present in the molecule. The neighborhood complexity indices developed by our group belong to the class of chemical complexity indices.

Calculation of Information Content (IC_r), Structural Information Content (SIC_r), and Complementary Information Content (CIC_r) Parameters

Definitions A graph $G=[V,E]$ consists of a finite set V whose elements are called vertices along with a set E of two element subsets of V, the elements of E being termed edges. The vertex set and edge set are denoted by $V(G)$ and $E(G)$, respectively. Two vertices in a graph are adjacent if they are connected by an edge. A multigraph is a graph having more than one edge between at least one pair of adjacent vertices. By the term graph, we mean a finite undirected graph (or multigraph) without self loops. A walk of length k in a graph G from vertex u_1 to vertex u_{k+1} is a sequence of vertices $u_1, u_2, \ldots, u_{k+1}$ for which the edge $(u_i, u_{i+1}) \in E(G)$ for $i = 1, 2, \ldots, k$. A walk is closed if $u_1 = u_{k+1}$, otherwise the walk is open. A path is an open walk in which all vertices are distinct. The distance $\partial(u, v)$ between the vertices $u, v \in V(G)$ is the length of the shortest path between u and v. The number of edges incident with a vertex u is called the degree of the vertex and is denoted by deg u. The eccentricity, $e(u)$, of a vertex u is defined as: $e(u) = \max_{u,v \in V(G)} \partial(u, v)$. The radius, ρ, of a graph is given by $\rho = \min_{u \in V(G)} e(u) = \min \max_{u,v \in V(G)} \partial(u, v)$. For a vertex $v \in V(G)$, the first-order neighborhood, $\Gamma^1(v)$ is a subset of $V(G)$ such that $\Gamma^1(v) = (u \in V(G) | \partial(u, v) = 1)$. The first-order closed neighborhood, $N^1(v)$, of v is defined as $N^1(v) = (v) U\Gamma^1(v) = \Gamma^0(v) U\Gamma^1(v)$ where (v) is the one point set consisting of v only and may be taken as $\Gamma^0(v)$. If ρ is the radius of a graph, one can construct $N^i(u)$, $i = 1, 2, \ldots, \rho$, for each vertex $u \in V(G)$.

The Basic Principle In the information-theoretic formalism an appropriate set A of n elements derived from the molecular graph is partitioned into h disjoint subsets A_i of order $n_i (i = 1, 2, \ldots, h; \sum_{i=1}^{h} n_i = n)$ by means of an equivalence relation defined on the set A. A probability scheme may then be attached to this distribution:

$$\begin{pmatrix} A_1, A_2, \ldots, A_h \\ p_1, p_2, \ldots, p_h \end{pmatrix}; \qquad \sum_i p_i = 1; \qquad p_i \geq 0 \quad (i = 1, 2, \ldots, h);$$

$$p_i = \frac{|A_i|}{|A|} = n_i/n, \quad (i = 1, 2, \ldots, h).$$

Here p_i is the probability that a randomly selected element of A will lie in the ith subset. The entropy (or complexity) of the structure associated with the particular mode of partitioning is then computed using Shannon's [22] formula:

$$\text{Information content} = -\sum p_i \log_2 p_i \text{ bits.} \tag{1}$$

It is apparent from the above that the measure of complexity of a structure depends both on the manner in which the set A is derived from the structure and the equivalence relation used to obtain the partition. For a given chemical structure the equivalence classes obtained from the decomposition of the vertex set of the hydrogen-suppressed graph will be different from the disjoint subsets derived from the vertex set of the total (nonhydrogen-suppressed) molecular graph. Rashevsky [26], Mowshowitz [29], and Truccó [30] calculated the information content of hydrogen-depleted graphs where topologically equivalent vertices (or vertices which constitute the orbits of the automorphism group) were placed in the same subset. Kier [31] calculated the information content of a total molecular graph where its vertex set was partitioned into equivalence classes on the basis of symmetry operations and experimental evidence from NMR spectra.

In the formalism developed by our group the total molecular graph was used to define various information-theoretic graph invariants and the method is sufficiently general to include linear graphs as well as multigraphs. Initially, the equivalence relation was defined on the vertex set $V(G)$ so that two vertices will be in a particular equivalence class if they have similar edge multiplicity and the same number of first order neighbors with similar degrees [20, 32]. Neither the chemical identity of the vertices nor the pattern of bonding of the higher order neighbors (elements situated at a distance of $2, 3, \ldots, \rho$ from a specified vertex) was considered. Subsequently Roy *et al.* [21] developed the formalism where various orders of information-theoretic graph invariants were calculated from the corresponding equivalence relation defined on $V(G)$.

The Equivalence Relation Let k be any non-negative integer, $0 \leq k \leq \rho$, ρ being the radius of the graph G. Two vertices u_0 and v_0 of G will be called equivalent with respect to the kth order neighborhood if and only if:

(i) $\deg u_0 = \deg v_0$, u_0 and v_0 being atoms of the same chemical element;
(ii) $|N^i(u_0)| = |N^i(v_0)|$ for $i = 1, 2, \ldots, k$, where $N_i(u_0)$ and $N_i(v_0)$ represent the cardinalities of $N^i(u_0)$ and $N^i(v_0)$, respectively; and
(iii) corresponding to each path $pu_0 = u_0, u_1, u_2, \ldots, u_\ell, u_\ell \in \Gamma^\ell(u_0)$ for $\ell = 1, 2, \ldots, k$, there is a corresponding path $pv_0 = v_0, v_1, v_2, \ldots, v_\ell$, $v_\ell \in \Gamma^\ell(v_0)$ for $\ell = 1, 2, \ldots, k$, such that: (a) deg u_i = deg v_i, u_i, and v_i being atoms of the same chemical element for $i = 1, 2, \ldots, k$, and

(b) $E(u_i, u_{i-1}) = E(v_i, v_{i-1})$, $i = 1, 2, \ldots, k$, where $E(u_i, u_{i-1})$ is the number of edges joining u_i with u_{i-1} while $E(v_i, v_{i-1})$ represents the number of edges connecting v_i and v_{i-1}.

The above relation is an equivalence relation by virtue of being reflexive, symmetric and transitive and is utilized to obtain the partitions of vertices of G with respect to different orders of neighborhoods.

The information content ($\mathrm{IC_k}$), which represents the information content per vertex of a graph with respect to the kth order neighborhood, is calculated using Eq. (1). The total information content (TIC_k), a measure of the complexity per graph, and other derived indices, structural information content (SIC_k), relative nonstructural information content (RNSIC_k), and complementary information content (CIC_k) may be computed as follows [33].

$$\mathrm{TIC}_k = n \times \mathrm{IC}_k \quad (2)$$

$$\mathrm{SIC}_k = \mathrm{IC}_k / \log_2 n \quad (3)$$

$$\mathrm{RNSIC}_k = (\log_2 n - \mathrm{IC}_k)/\log_2 n \quad (4)$$

$$\mathrm{CIC}_k = \log_2 n - \mathrm{IC}_k \quad (5)$$

where n is the cardinality of $V(G)$.

Mowshowitz [29] has pointed out that the information content of graphs may be looked upon as quantitative measures of their relative complexity. For a particular order of neighborhood, IC_k is maximum when all the vertices are distinct. On the other hand, for the most symmetric graph (where all the vertices are equivalent) IC_k is equal to zero. CIC_k is equal to $\log_2 n$ when a graph is most symmetric with respect to kth order neighborhood and CIC_k vanishes when all the vertices are distinct. Since the above indices are derived from the consideration of neighborhoods of the vertices of chemical graphs, they are also called indices of neighborhood symmetry.

The first and second order neighborhoods of vertices of the labeled graph of 1-butanol are given in Figures 2 and 3, respectively.

For the first order neighborhood, the $\mathrm{IC_1}$, $\mathrm{SIC_1}$ and $\mathrm{CIC_1}$ indices for 1-butanol are calculated as:

$$\mathrm{IC}_1 = 5 * 1/12 \log_2 1/12 + 7/12 \log_2 7/12 = 1.950 \text{ bits} \quad (6)$$

$$\mathrm{SIC}_1 = \mathrm{IC}_1/\log_2 12 = 0.544 \quad (7)$$

$$\mathrm{CIC}_1 = \log_2 12 - \mathrm{IC}_1 - 1.635 \text{ bits.} \quad (8)$$

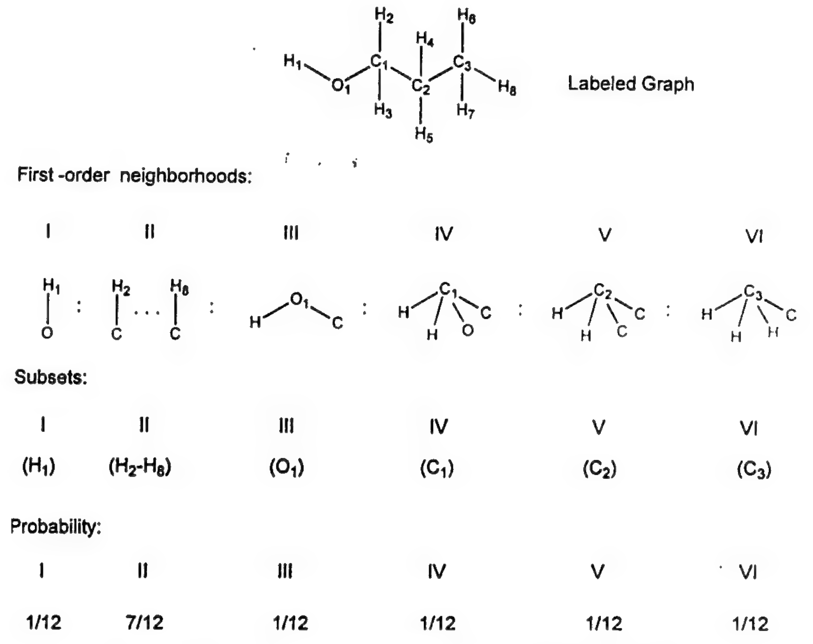


Figure 2 First order neighborhoods for 1-butanol.

For the second order neighborhood, the indices for 1-butanol are:

$$\mathrm{IC}_2 = 5 * 1/12 \log_2 12 + 2 * 2/12 \log_2 12/2 + 3/12 \log_2 12/3 = 2.855 \text{ bits} \quad (9)$$

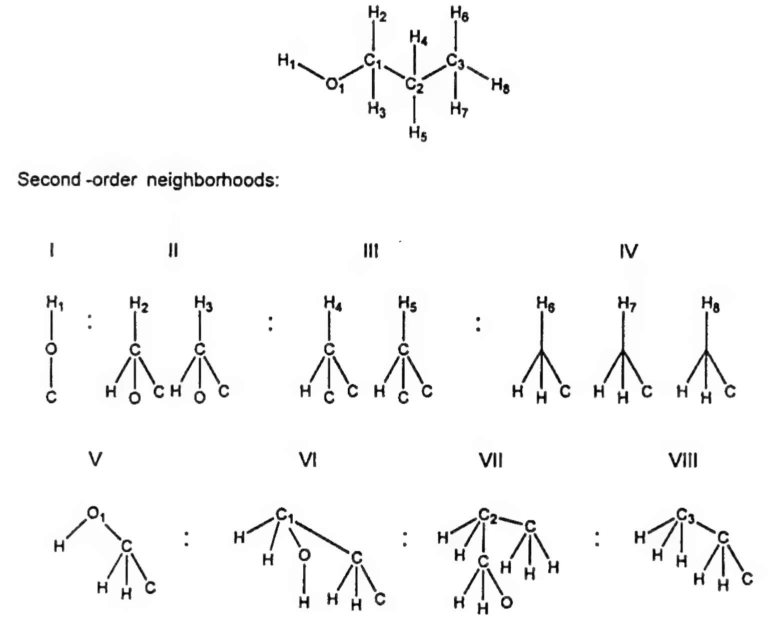
$$\mathrm{SIC}_2 = \mathrm{IC}_2/\log_2 12 = 0.796 \quad (10)$$

$$\mathrm{SIC}_2 = \log_2 12 - \mathrm{IC}_2 = 0.730. \quad (11)$$

APPLICATIONS OF NEIGHBORHOOD COMPLEXITY PARAMETERS

The Nature of Neighborhood Complexity Indices and Their Intercorrelation with Other Topological Indices

The formation theoretic indices of various orders have been calculated and used in QSAR studies by our group [4–14]. One interesting point was to



Subsets:

I	II	III	IV	V	VI	VII	VIII
(H_1)	(H_2-H_3)	(H_4-H_5)	(H_6-H_8)	(O_1)	(C_1)	(C_2)	(C_3)

Probability:

I	II	III	IV	V	VI	VII	VIII
1/12	2/12	2/12	3/12	1/12	1/12	1/12	1/12

Figure 3 Second order neighborhoods for 1-butanol.

investigate how far these indices are independent of (uncorrelated with) other graph invariants. Basak *et al.* [34, 35] studied the intercorrelation of a set of ninety TIs calculated for a very diverse set of 3692 chemicals. It was found that the lower order IC, SIC, and CIC indices are strongly correlated with connectivity indices, the Wiener index *W* and molecular size. But the higher order information indices are not correlated with any of the most well known graph invariants. In the principal components analysis (PCA) with the ninety parameters mentioned above, the higher order neighborhood complexity parameters were strongly correlated with the second principal component (PC_2) as shown in Table I.

Characterization of Structure using Neighborhood Complexity Parameters

An important property of any molecular descriptor is its discriminatory power. So, it was of interest to see how far the neighborhood complexity indices can discriminate closely related structures. To this end, we attempted to discriminate a set of thirty eight structures consisting of nineteen pairs of isospectral graphs (Figure 4). Table II gives the values of three connectivity indices and the first, second, and third order IC indices. It is evident from the data that the information theoretic parameters have a reasonably good discriminatory power for the difficult set of isospectral graphs [1].

QUANTITATIVE STRUCTURE–ACTIVITY/PROPERTY RELATIONSHIP STUDIES USING COMPLEXITY INDICES

Toxicity of Monoketones

Monoketones are a group of industrial solvents. So, it was of interest to see whether we could correlate their toxicity using information theoretic molecular descriptors. Basak *et al.* [36] attempted to correlate the toxicity (LD_{50}) of a set of thirteen monoketones using information theoretic parameters. The result is summarized in Table III and IV. It is clear from the data that the LD_{50} values of the sets of monoketones can be predicted reasonably well from the calculated complexity indices.

Aquatic Toxicity of Alcohols

Aliphatic alcohols are a group of industrial chemicals which bring about their characteristic toxic effects by nonspecific and reversible action on protoplasmic structures of the cell. Roy *et al.* [21] successfully used neighborhood complexity indices to correlate the aquatic toxicity (LC_{50}) of ten alcohols in *Pimephales promelas* (fathead minnow), the data being

Table I Correlation coefficients of variables with the principal components (only the 10 most highly correlated are listed)

PC1		*PC2*		*PC3*		*PC4*		*PC5*		*PC6*		*PC7*		*PC8*		*PC9*		*PC10*	
K_1	0.959	SIC_3	0.973	${}^{4}\chi^{b}_{C}$	0.694	${}^{4}\chi_{CH}$	0.848	${}^{6}\chi_{CH}$	−0.465	IC_0	0.501	${}^{4}\chi^{v}_{C}$	0.538	K_{10}	−0.319	${}^{5}\chi^{v}_{PC}$	0.292	IC_0	0.282
${}^{2}\chi$	0.954	CIC_4	−0.955	${}^{3}\chi^{b}_{C}$	0.693	${}^{4}\chi^{b}_{CH}$	0.844	${}^{6}\chi^{v}_{CH}$	−0.457	SIC_0	0.424	${}^{3}\chi^{v}_{C}$	0.494	${}^{6}\chi^{v}$	−0.311	${}^{6}\chi^{v}_{PC}$	0.285	${}^{5}\chi^{v}_{C}$	0.282
${}^{3}\chi$	0.954	CIC_3	−0.952	${}^{5}\chi^{b}_{C}$	0.683	${}^{4}\chi^{v}_{CH}$	0.795	${}^{6}\chi^{b}_{CH}$	0.437	${}^{6}\chi^{v}_{CH}$	−0.374	${}^{6}\chi^{b}_{C}$	−0.480	${}^{5}\chi^{v}_{C}$	−0.309	${}^{4}\chi^{v}_{PC}$	0.282	${}^{3}\chi^{v}$	0.272
K_2	0.953	SIC_4	0.947	${}^{4}\chi_{C}$	0.680	${}^{3}\chi_{CH}$	0.751	${}^{3}\chi^{b}_{CH}$	0.406	O	−0.349	${}^{6}\chi_{C}$	−0.434	K_9	−0.301	${}^{4}\chi_{C}$	0.273	${}^{3}\chi^{b}_{C}$	−0.233
K_0	0.949	SIC_2	0.940	${}^{3}\chi^{v}_{C}$	0.668	${}^{3}\chi^{b}_{CH}$	0.751	${}^{3}\chi_{CH}$	0.406	SIC_0	0.334	${}^{5}\chi^{b}_{C}$	−0.391	${}^{5}\chi_{CH}$	0.289	K_9	0.268	K_9	−0.232
${}^{1}\chi$	0.942	CIC_5	−0.940	${}^{5}\chi_{C}$	0.644	${}^{3}\chi^{b}_{CH}$	0.740	${}^{3}\chi^{v}_{CH}$	0.391	${}^{3}\chi^{b}_{C}$	0.318	${}^{5}\chi_{C}$	−0.343	${}^{5}\chi^{b}_{CH}$	0.287	K_{10}	0.264	K_8	−0.230
${}^{3}\chi^{b}$	0.938	CIC_6	−0.922	${}^{6}\chi_{C}$	0.637	${}^{3}\chi^{v}_{CH}$	0.718	${}^{4}\chi^{b}_{C}$	0.316	${}^{3}\chi^{v}_{CH}$	0.314	${}^{2}\chi^{v}$	0.304	${}^{6}\chi^{v}_{C}$	−0.281	SIC_0	0.249	${}^{1}\chi^{v}$	0.224
${}^{4}\chi$	0.935	SIC_5	0.915	${}^{3}\chi_{C}$	0.612	${}^{5}\chi_{CH}$	0.707	${}^{5}\chi^{v}_{CH}$	−0.311	${}^{3}\chi_{CH}$	0.314	${}^{4}\chi_{C}$	0.274	${}^{5}\chi^{v}_{C}$	0.277	${}^{5}\chi_{CH}$	0.239	SIC_0	0.222
${}^{4}\chi^{b}$	0.934	SIC_6	0.887	${}^{6}\chi^{b}_{C}$	0.602	${}^{5}\chi^{v}_{CH}$	0.672	${}^{6}\chi^{b}_{C}$	0.304	${}^{3}\chi^{b}_{CH}$	0.314	${}^{4}\chi^{b}_{C}$	0.232	${}^{5}\chi^{v}$	−0.243	IC_0	0.235	${}^{2}\chi^{v}$	0.218
${}^{0}\chi$	0.934	CIC_2	−0.869	${}^{6}\chi^{v}_{C}$	0.600	${}^{6}\chi^{b}_{CH}$	0.472	${}^{6}\chi_{C}$	0.310	CIC_1	0.312	${}^{3}\chi^{b}_{C}$	0.210	K_8	−0.228	K_8	0.228	${}^{6}\chi^{v}_{C}$	0.212

catalogued in Table V. The following significant equations were derived for this set of alcohols:

$$\log \mathrm{LC}_{50} = 1.98 - 1.90\,\mathrm{CIC}_1 \qquad n = 10, \quad r = 0.99, \quad s = 0.32 \tag{12}$$

$$\log \mathrm{LC}_{50} = -33.8 + 27.4\,\mathrm{IC}_0 \qquad n = 10, \quad r = 0.99, \quad s = 0.35 \tag{13}$$

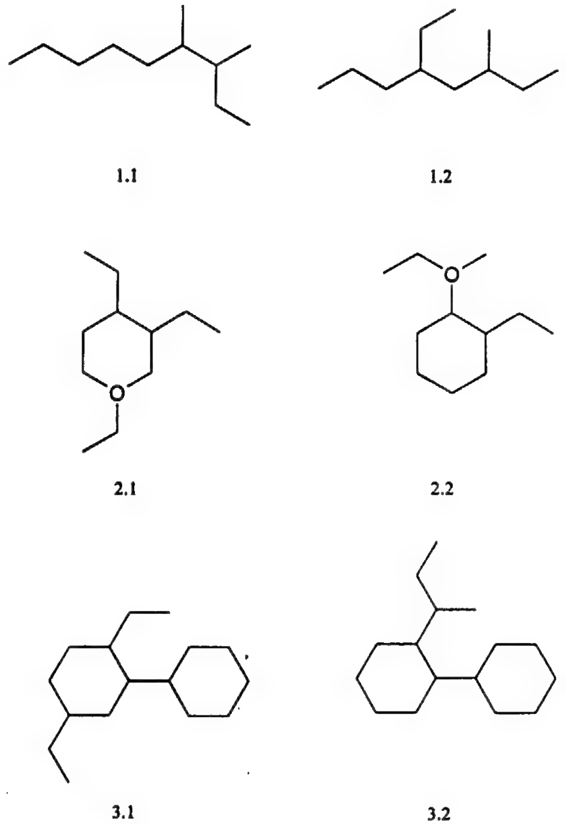


Figure 4 Thirty-eight isospectral graphs.

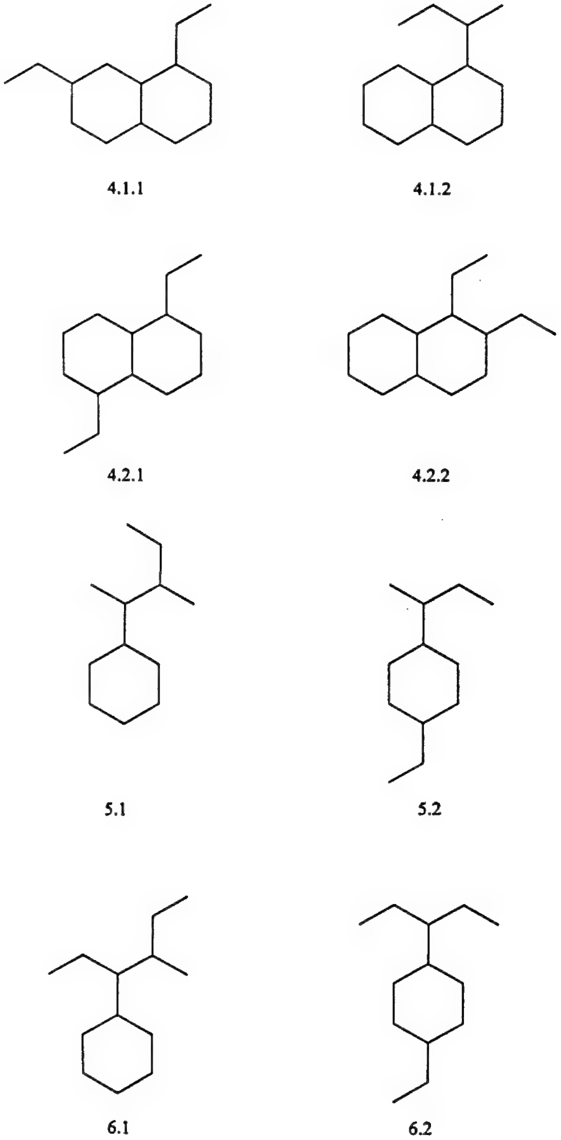


Figure 4 (*Continued*).

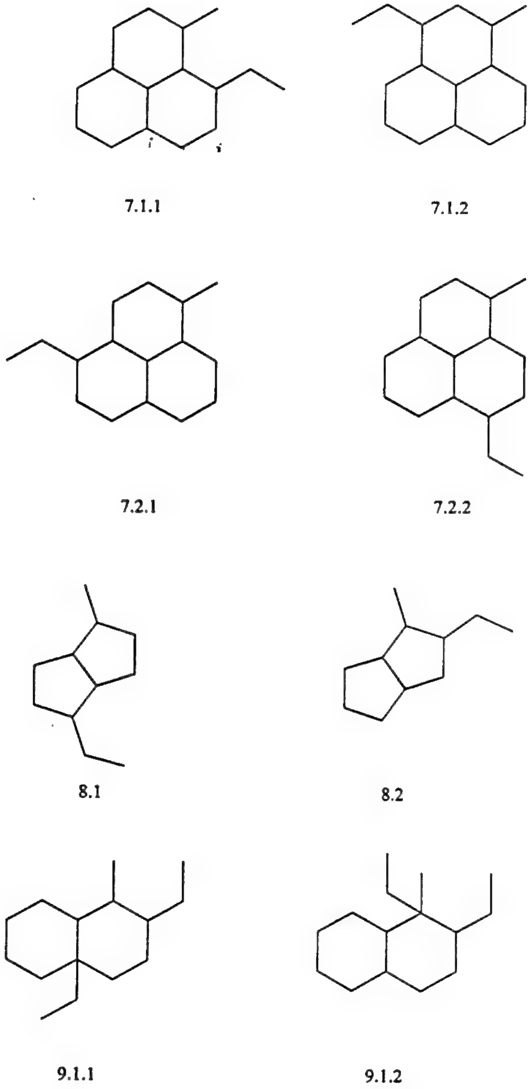


Figure 4 (*Continued*).

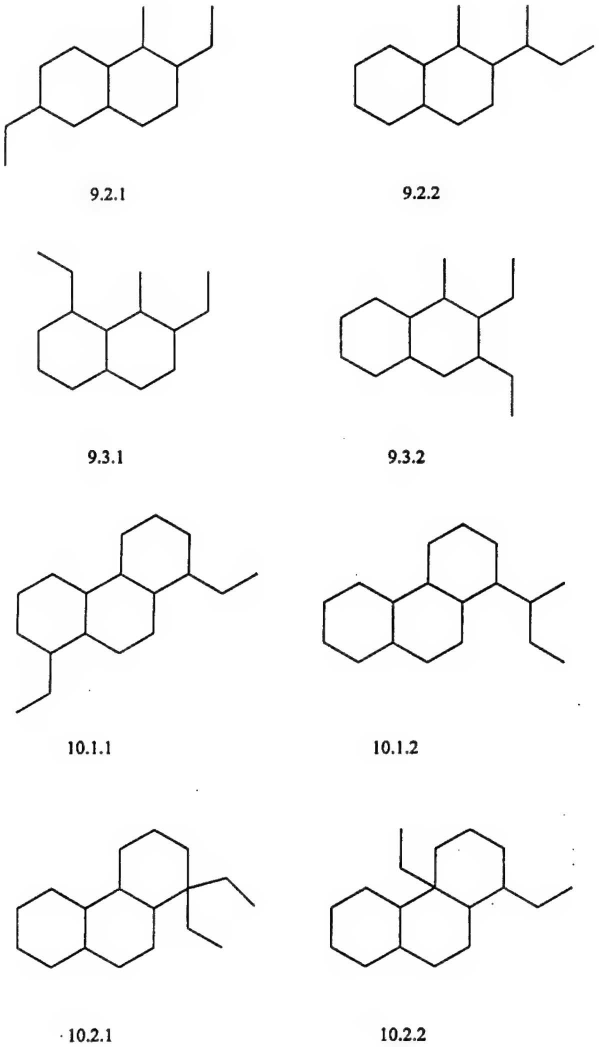


Figure 4 (*Continued*).

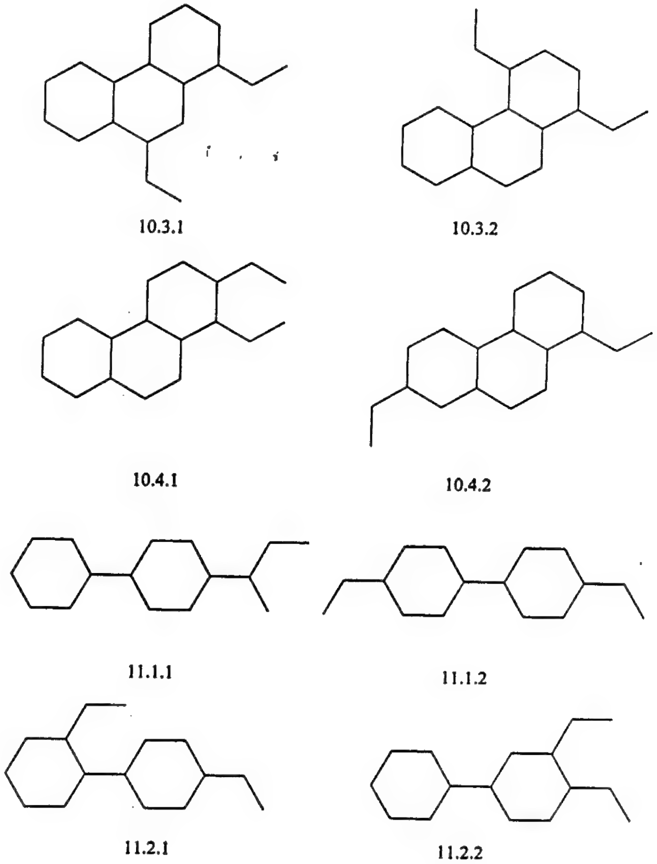


Figure 4 (*Continued*).

Mutagenicity of Nitrosamines

Nitrosamines are an important class of chemicals as many of them are carcinogens and mutagens. Basak *et al.* [37] correlated the mutagenic potency, ln R (R being the number of revertants per nanomole of the chemical in the Ames' test), of a group of 15 nitrosamines using IC_0 and IC_1 indices. The data for the mutagenicity and the values of IC parameters for the fifteen mutagens are given in Table VI. The following regression

Table II Selected topological indices for 38 isospectral graphs (Figure 4)

Graph	$^0\chi$	$^1\chi$	$^2\chi$	IC_0	IC_1	IC_2
1.1	8.690	5.219	3.859	0.898	1.368	2.665
1.2	8.690	5.240	3.812	0.898	1.368	2.701
2.1	8.975	5.812	4.424	0.918	1.418	2.675
2.2	8.975	5.791	4.502	0.918	1.418	2.828
3.1	11.380	7.847	6.318	0.932	1.384	2.726
3.2	11.380	7.826	6.396	0.932	1.384	2.664
4.1.1	9.966	6.847	5.610	0.934	1.417	2.784
4.1.2	9.966	6.826	5.689	0.934	1.417	2.765
4.2.1	9.966	6.864	5.526	0.934	1.417	2.684
4.2.2	9.966	6.864	5.526	0.934	1.417	2.684
5.1	8.975	5.753	4.643	0.918	1.418	2.807
5.2	8.975	5.774	4.575	0.918	1.418	2.717
6.1	9.682	6.291	4.856	0.918	1.404	2.789
6.2	9.682	6.312	4.766	0.918	1.404	2.565
7.1.1	11.121	7.809	6.906	0.946	1.457	2.794
7.1.2	11.121	7.809	6.908	0.946	1.457	2.982
7.2.1	11.121	7.809	6.896	0.946	1.457	2.856
7.2.2	11.121	7.809	6.896	0.946	1.457	2.856
8.1	7.845	5.326	4.628	0.938	1.469	2.802
8.2	7.845	5.326	4.618	0.938	1.469	2.995
9.1.1	10.889	7.232	6.134	0.933	1.517	2.978
9.1.2	10.889	7.220	6.193	0.933	1.517	2.885
9.2.1	10.836	7.258	6.116	0.933	1.458	2.928
9.2.2	10.836	7.236	6.194	0.933	1.458	2.928
9.3.1	10.836	7.274	6.041	0.933	1.458	2.864
9.3.2	10.836	7.274	6.004	0.933	1.458	2.974
10.1.1	12.535	8.847	7.431	0.943	1.429	2.664
10.1.2	12.535	8.809	7.594	0.943	1.429	2.729
10.2.1	12.588	8.805	7.518	0.943	1.483	2.764
10.2.2	12.588	8.815	7.482	0.943	1.483	2.764
10.3.1	12.535	8.847	7.443	0.943	1.429	2.760
10.3.2	12.535	8.847	7.441	0.943	1.429	2.729
10.4.1	12.535	8.847	7.431	0.943	1.429	2.664
10.4.2	12.535	8.830	7.516	0.943	1.429	2.769
11.1.1	11.380	7.809	6.458	0.932	1.384	2.589
11.1.2	11.380	7.830	6.378	0.932	1.384	2.438
11.2.1	11.380	7.847	6.306	0.932	1.384	2.622
11.2.2	11.380	7.847	6.308	0.932	1.384	2.595

Table III Oral LD_{50}, log P, and topological indices for monoketones

Compound	*Control* LD_{50}*	CCl_4 LD_{50}*	log P	TIC_0	TIC_1	CIC_0	CIC_1
Acetone	90.39	73.35	− 0.48	12.955	15.710	2.026	1.751
Methyl ethyl ketone	56.16	45.86	0.26	16.106	22.108	2.462	2.000
Methyl *n*-propyl ketone	25.60	23.13	0.78	19.171	26.781	2.802	2.326
Methyl isopropyl ketone	29.86	26.98	0.56	19.171	26.026	2.802	2.373
Methyl *n*-butyl ketone	24.26	16.17	1.19	22.181	30.936	3.080	2.620
Methyl isobutyl ketone	26.66	19.75	1.31	22.181	32.936	3.080	2.514
Methyl *n*-amyl ketone	21.08	10.39	2.03	25.153	34.804	3.316	2.877
Methyl isoamyl ketone	22.26	10.99	1.88	25.153	38.050	3.316	2.730
Methyl *n*-hexyl ketone	29.82	12.38	2.37	28.096	38.487	3.520	3.104
Methyl *n*-heptyl ketone	56.19	26.59	3.14	31.018	42.038	3.700	3.306
Methyl 3-methylhexyl ketone	33.80	23.90	2.92	31.018	46.792	3.700	3.136
Methyl *n*-octyl ketone	50.79	15.99	3.73	33.922	45.790	3.860	3.487
Methyl *n*-nonyl ketone	114.40	32.07	4.09	36.812	48.866	4.005	3.650

*Expressed as mmole/kg.

equations were developed on this data set:

$$\ln R = 61.00 - 86.80(\mathrm{IC}_0) + 29.20(\mathrm{IC}_0)^2 \qquad (14)$$
$$n = 15, \quad r = 0.96, \quad s = 1.17, \quad p < 0.001$$

$$\ln R = 12.00 - 15.30(\mathrm{IC}_1) + 3.84(\mathrm{IC}_1)^2 \qquad (15)$$
$$n = 15, \quad r = 0.98, \quad s = 0.86, \quad p < 0.001.$$

Binding of Barbiturates to Cytochrome P_{450}

Cytochrome P_{450} is a microsomal enzyme which is involved in the metabolism of drugs. Hepatic microsomal enzymes play an important role in the metabolism of drugs and xenobiotics. The initial step in the metabolic process is the binding of the chemical by the membrane. So, it was of interest to see whether we could correlate the binding affinity (K_s)

Table IV Parabolic correlation of LD_{50} values with log P and four topological indices

Independent Variable (X)	*LD_{50} (Control) = $A + BX + CX^2$*						*LD_{50} (CCl_4) = $A + BX + CX^2$*					
	A	*B*	*C*	r^a	*SD*	*F*	*A*	*B*	*C*	r^2	*SD*	*F*
log *P*	62.20	−49.70	14.30	0.94	11.04	35.94	50.50	−34.00	7.34	0.94	6.70	34.82
TIC_0	340.00	−26.40	0.54	0.96	9.13	54.87	216.00	−15.00	0.28	0.95	6.10	43.12
TIC_1	288.00	−16.30	0.25	0.86	16.10	14.25	195.00	−9.85	0.13	0.97	4.68	76.61
CIC_0	718.00	−457.00	74.80	0.91	12.99	24.57	407.00	−235.00	35.10	0.97	4.76	74.05
CIC_1	620.00	−448.00	83.50	0.95	9.62	48.88	364.00	−239.00	40.70	0.96	5.54	53.27

[a] For each equation, *r* is the correlation coefficient, *SD* the standard deviation, and *F* the *F*-ratio between the variances of observed and calculated values.

Table V Aquatic toxicity and information indices for aliphatic alcohols

Name	log LC_{50}*	CIC_1	IC_0
Methanol	−0.06	0.79	1.25
Ethanol	−0.51	1.29	1.22
2-Propanol	−0.80	1.80	1.19
1-Butanol	−1.63	2.04	1.16
1-Hexanol	−3.02	2.67	1.12
1-Octanol	−4.00	3.15	1.09
1-Nonanol	−4.40	3.35	1.07
1-Decanol	−4.84	3.52	1.06
1-Undecanol	−5.22	3.68	1.06
1-Dodecanol	−5.27	3.83	1.05

*Expressed as mole/litre.

Table VI Mutagenicity and topological parameters for nitrosamines

No.	*Compound*	ln R^a	IC_0	IC_1
1	Dipropyl-*N*-nitrosamine	−2.53	1.444	1.945
2	Dibutyl-*N*-nitrosamine	−1.90	1.373	1.856
3	Dipentyl-*N*-nitrosamine	−3.00	1.320	1.769
4	*N*-Nitrosopyrrolidine	−3.91	1.640	2.040
5	*N*-Nitrosomorpholine	−2.81	1.750	2.250
6	*N*-Nitrosopiperidine	−4.60	1.568	1.949
7	*N*-Methyl-*N*-nitroso-*N'*-nitroguanidine	7.23	2.108	3.578
8	*N*-Ethyl-*N*-nitroso-*N'*-nitroguanidine	5.86	2.063	3.532
9	*N*-Propyl-*N*-nitroso-*N'*-nitroguanidine	3.69	2.006	3.475
10	*N*-Butyl-*N*-nitroso-*N'*-nitroguanidine	3.89	1.948	3.343
11	*N*-Isobutyl-*N*-nitroso-*N'*-nitroguanidine	4.34	1.948	3.343
12	*N*-Pentyl-*N*-nitroso-*N'*-nitroguanidine	3.09	1.894	3.207
13	*N*-Hexyl-*N*-nitroso-*N'*-nitroguanidine	1.67	1.844	3.080
14	*N*-Nitrosomethylurea	1.48	1.888	3.022
15	*N*-Nitrosoethylurea	0.10	1.830	3.000

[a] Natural logarithm of revertants per nanomole, determined by the Ames' mutagenicity assay.

Table VII Topological parameters, log P, and K_s values for 5-ethyl-5-alkyl barbiturates

No.	*Alkyl Group*	K_s [a]	log P [a]	H^D	IC_0	SIC_0
1	Propyl	0.235	0.87	3.58	1.64	0.34
2	Butyl	0.089	1.70	3.67	1.60	0.32
3	Pentyl	0.032	2.23	3.75	1.56	0.31
4	3-Methylbutyl	0.038	2.11	3.76	1.56	0.31
5	1-Methylbutyl	0.045	2.13	3.78	1.56	0.31
6	2,3-Dimethylbutyl	0.025	2.39	3.86	1.53	0.29
7	Hexyl	0.019	3.08	3.82	1.53	0.29
8	Heptyl	0.020	3.64	3.89	1.50	0.28
9	Octyl	0.024	3.85	3.96	1.47	0.27
10	Nonyl	0.056	4.13	4.02	1.45	0.26

[a] Experimental K_s (mM).

of molecules from their theoretical molecular descriptors. Basak [38] correlated the K_s values of a set of ten barbiturates using lipophilicity (log P) and three information theoretic parameters, the data being given in Table VII.

$$K_s = 0.45 - 0.29(\log P) + 0.05(\log P)^2 \quad (16)$$
$$n = 10, \quad r = 0.99, \quad s = 0.01, \quad F_{2,7} = 313.45$$

$$K_s = 27.79 - 36.78(\mathrm{IC}_0) + 12.17(\mathrm{IC}_0)^2 \quad (17)$$
$$n = 10, \quad r = 0.99, \quad s = 0.01, \quad F_{2,7} = 156.14$$

$$K_s = 5.94 - 41.26(\mathrm{SIC}_0) + 71.84(\mathrm{SIC}_0)^2 \quad (18)$$
$$n = 10, \quad r = 0.99, \quad s = 0.01, \quad F_{2,7} = 224.34$$

Anesthetic Potency and Toxicity of Barbiturates

Barbiturates constitute a group of chemicals which act on the biological system by nonspecific narcotic mechanisms. It is usually known that lipophilicity is the parameter of choice for predicting narcotic action of chemicals [39]. Basak *et al.* [40] carried out a comparative study of log P, connectivity indices and a few information theoretic indices to correlate the anesthetic potency (AD_{50}) of a set of thirteen barbiturate derivatives (Table VIII). The result of regression analysis is given in Table IX.

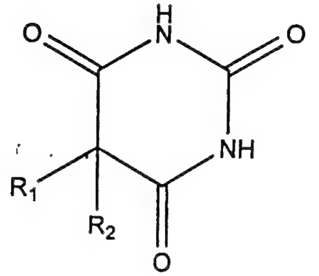


Figure 5 Hydrogen-suppressed graph of barbiturate.

A similar comparative study of log P and TIs was carried out to predict the inhibition of *Arbacia* egg cell division by a group of twenty three barbiturates (Table X). The result of regression analysis of this set of compounds is given in Table XI.

DISCUSSION

The objectives of this chapter were: (a) to provide to the reader the theoretical basis of the information theoretic indices of neighborhood complexity, and (b) to review the utility of these parameters in the characterization of structures and prediction of properties.

Various topological indices are being used by researchers in QSAR/QSPR studies. These parameters will be used more and more in predictive pharmacology and toxicology because such indices are algorithmically derived, i.e., they can be computed without any error using computer software. But we need to know which of the large number of TIs encode unique and non-redundant information. One solution to this problem is to use orthogonal variables derived from TIs using techniques like principal components analysis (PCA). Basak *et al.* [34] carried out PCA of a diverse set of 3692 chemicals and found that first four PCs explained about 80% of the variance in the data and the first ten PCs with eigenvalues greater than or equal to 1.0 explained about 93% of variance in the data. Another important finding was that while the first PC was strongly correlated with the Wiener index, I_D^W, $\bar{I}_D^W$, connectivity indices and the lower order neighborhood complexity parameters, the higher order neighborhood indices were correlated with the second PC which was minimally correlated with other known topological indices. This has been found to be true for

Table VIII Lipophilicity, anesthetic dose (AD_{50}) in mice, and molecular descriptors for barbiturates (Figure 5)

No.	R_1	R_2	AD_{50}	$\log P$	TIC_0	TIC_1	CIC_0	W	I_D^W	$\bar{I}_D^W$	$^1\chi$	$^1\chi^v$
1	Methyl	1-Methyl, 1-Propenyl	2.64	0.65	43.90	73.48	3.01	272	1727	6.35	6.42	4.22
2	Ethyl	1-Methyl, 1-Propenyl	3.15	1.15	47.60	82.35	3.21	324	2127	6.56	6.98	4.78
3	Propyl	1-Methyl, 1-Propenyl	3.29	1.65	51.19	89.18	3.40	391	2640	6.75	7.48	5.28
4	Allyl	1-Methyl, 1-Propenyl	3.39	1.35	49.09	89.43	3.27	391	2640	6.74	7.48	4.89
5	Butyl	1-Methylvinyl	3.36	2.15	54.69	95.26	3.56	474	3281	6.92	7.98	5.78
6	Methyl	1-Methylvinyl	2.12	0.15	40.06	67.28	2.78	218	1338	6.14	5.88	3.71
7	Ethyl	1-Methylvinyl	2.91	0.65	43.90	76.23	3.01	265	1687	6.36	6.44	4.27
8	Propyl	1-Methylvinyl	3.04	1.15	47.60	83.10	3.21	326	2139	6.56	6.94	4.77
9	Allyl	1-Methylvinyl	3.06	0.85	45.38	82.40	3.07	326	2139	6.56	6.91	4.38
10	Butyl	1-Methylvinyl	3.33	1.65	51.19	89.18	3.40	402	2709	6.74	7.44	5.27
11	Isobutyl	1-Methylvinyl	3.27	1.45	51.19	91.18	3.40	389	2627	6.75	7.30	5.13
12	Amyl	1-Methylvinyl	3.32	2.15	54.69	94.77	3.56	494	3409	6.90	7.94	5.77
13	Isoamyl	1-Methylvinyl	3.26	1.95	54.69	98.01	3.56	480	3318	6.91	7.80	5.63

Table IX Correlation of AD_{50} with $\log P$ and topological indices for barbiturates (Figure 5) ($AD_{50} = a + bx + cx^2$)

X	a	b	c	n	r	s	F
TIC_0	−0.139	0.660E−1	0	13	0.85	0.20	29.38
	−18.50	0.833	−0.796E−2	13	0.97	0.10	80.02
TIC_1	0.522E−1	0.355E−1	0	13	0.90	0.16	46.49
	−12.10	0.330	−0.177E−2	13	0.99	0.06	196.98
CIC_0	−1.03	1.26	0	13	0.86	0.19	32.25
	−27.80	18.00	−2.60	13	0.97	0.10	69.67
W	1.84	0.342E−2	0	13	0.82	0.21	23.00
	−1.37	0.220E−1	−0.256E−4	13	0.97	0.10	74.42
I_D^W	1.99	0.450E−3	0	13	0.82	0.22	22.61
	−0.530	0.266E−2	−0.453E−6	13	0.97	0.10	75.14
$\bar{I}_D^W$	−5.96	1.36	0	13	0.91	0.15	53.98
	−104.00	31.30	−2.28	13	0.98	0.08	130.37
$^1\chi$	−0.584	0.513	0	13	0.90	0.16	48.02
	−17.20	5.28	−0.340	13	0.98	0.08	131.38
$^1\chi^v$	0.757	0.474	0	13	0.85	0.20	27.73
	−8.74	4.46	−0.412	13	0.96	0.10	67.12
$\log P$	2.44	0.498	0	13	0.85	0.20	29.13
	1.93	1.58	−0.438	13	0.97	0.10	76.40

small databases of congeneric chemicals like hydrocarbons as well as diverse and the largest set of chemicals analyzed by us ($n > 40{,}000$). So, one can say that the higher order neighborhood indices encode structural information not quantified by any other class of TIs.

Another important property of any structural descriptor is its discriminatory power. We tested this aspect of IC, SIC, and CIC indices using a set of nineteen pairs of isospectral graphs. A comparison of connectivity indices and our information theoretic indices shows that the higher order information indices and connectivity indices have reasonably good ability of discriminating the set of thirty eight graphs [1]. In some cases, e.g., graphs 10.3.1 and 10.3.2, the values of the $^2\chi$ index are almost identical whereas the IC_2 index is slightly more discriminating (Table II).

One practical use of TIs is in the development of QSAR models for predictive toxicology and pharmacology. In this chapter we have reviewed the results of our QSPR/QSAR studies of various congeneric sets of

Table X Lipophilicity, log (1/C), the potency for inhibition of *Arbacia* egg cell division, and molecular descriptors for barbiturates (Figure 5)

No.	R_1	R_2	log (1/C)	log *P*	TIC_0	TIC_1	CIC_0	*W*	I_D^W	$\bar{I}_D^W$	$^1\chi$	$^1\chi^v$
1	Ethyl	2-Ethylhexyl	3.70	3.45	63.23	104.20	3.95	690	4984	7.22	8.99	7.16
2	Allyl	1-Methylbutyl	3.62	2.15	54.69	98.01	3.56	472	3268	6.92	7.98	5.76
3	Benzyl	Isopropyl	3.40	2.64	55.46	101.85	3.54	644	4660	7.23	8.96	6.17
4	Allyl	Benzyl	2.82	2.54	53.10	95.18	3.43	660	4773	7.23	9.08	5.90
5	Ethyl	1-Methyl-2-Butenyl	3.30	1.65	51.10	89.18	3.40	398	2683	6.74	7.48	5.28
6	Ethyl	Hexyl	3.12	2.65	56.53	89.42	3.68	532	3655	6.87	8.06	6.23
7	2-Methylallyl	1-Methylbutyl	3.17	2.45	58.12	104.09	3.71	548	3888	7.09	8.34	6.15
8	Ethyl	Isoamyl	2.82	1.95	53.10	88.70	3.52	418	2811	6.72	7.41	5.59
9	Ethyl	1-Methylbutyl	2.92	1.95	53.10	88.70	3.52	398	2683	6.74	7.48	5.65
10	Ethyl	1-Ethylpropyl	2.85	1.95	53.10	88.70	3.52	387	2615	6.75	7.52	5.69
11	Ethyl	Amyl	2.82	2.15	53.10	84.60	3.52	431	2892	6.71	7.56	5.73
12	Ethyl	2-Phenylethyl	2.66	2.80	55.46	99.10	3.54	700	5044	7.20	9.08	6.29
13	Ethyl	1,3-Dimethylbutyl	2.82	2.25	56.53	94.93	3.68	474	3278	6.91	7.84	6.01
14	Ethyl	Cyclopentyl	2.77	0.79	49.09	85.43	3.27	382	2578	6.75	7.60	5.47
15	Allyl	Isobutyl	2.41	1.65	51.19	91.94	3.40	402	2711	6.74	7.42	5.19
16	Ethyl	Cyclohexenyl	2.24	1.20	52.69	90.79	3.44	458	3169	6.92	8.10	5.97
18	Ethyl	Butyl	2.40	1.65	49.60	79.52	3.35	346	2262	6.54	7.06	5.23
19	Ethyl	Phenyl	2.02	1.42	48.08	85.29	3.20	458	3169	6.92	8.10	5.33
20	Allyl	Isopropyl	2.01	1.15	47.60	85.10	3.21	326	2139	6.56	6.94	4.72
21	Allyl	Allyl	1.79	1.05	45.38	78.40	3.07	338	2215	6.55	7.06	4.45
22	Ethyl	Isopropyl	1.79	0.95	46.01	76.07	3.16	265	1687	6.36	6.44	4.61
23	Ethyl	Ethyl	1.49	0.65	42.32	68.12	2.95	220	1351	6.14	6.06	4.23

Table XI Correlation of log (1/C) with log *P* and topological indices for barbiturates (Figure 5) log (1/C) = $a + bx + cx^2$

X	*a*	*b*	*c*	*n*	*R*	*s*	*F*
TIC_0	−2.93	0.107	0	23	0.83	0.33	48.25
	−9.22	0.350	−0.232E−2	23	0.84	0.33	24.68
TIC_1	−1.81	0.498E−1	0	23	0.77	0.39	29.66
	−5.28	0.130	−0.452E−3	23	0.77	0.40	14.50
CIC_0	−4.82	2.17	0	23	0.83	0.34	47.63
	−14.70	7.94	−0.841	23	0.84	0.34	23.84
W	1.30	0.302E−1	0	23	0.65	0.46	15.75
	−0.803	0.124E−1	−0.978E−5	23	0.73	0.43	11.23
I_D^W	1.47	0.383E−3	0	23	0.65	0.46	15.20
	−0.262	0.151E−2	0.168E−6	23	0.72	0.43	11.06
$\bar{I}_D^W$	−7.25	1.45	0	23	0.68	0.44	18.56
	−51.60	14.60	−0.970	23	0.71	0.44	9.93
$^1\chi$	−1.13	0.489	0	23	0.66	0.46	15.86
	−12.80	3.53	−0.196	23	0.71	0.44	10.08
$^1\chi^v$	−1.30	0.710	0	23	0.80	0.36	37.21
	−4.92	2.03	−0.119	23	0.81	0.36	19.07
log *P*	1.46	0.646	0	23	0.78	0.38	31.69
	1.18	0.977	−0.864E−1	23	0.78	0.39	15.57

chemicals. It is clear that the neighborhood indices were able to predict biochemical, pharmacological, and toxicological properties of various congeneric sets of chemicals. We also used neighborhood complexity parameters in hierarchical QSAR analysis. The goal of hierarchical analysis is to use non-redundant and progressively more complex parameters in the development of QSAR models. Topostructural, topochemical, geometrical as well as semiempirical quantum chemical indices were used in such studies. The result of such QSAR studies showed that information theoretic indices encode structural information not quantified by other TIs and such indices are useful in the development of models for the prediction of different physicochemical, biomedicinal, and toxicological properties of molecules [41–44].

In conclusion, the neighborhood complexity parameters developed by us contain some unique structural information not coded by other topological indices. These indices, particularly, the higher order parameters, have sufficient discriminatory power so that they can be useful in the characterization of closely related structures. Finally, QSAR studies of various groups of chemicals show that the IC, SIC, and CIC indices are useful descriptors for QSPR/QSAR studies.

Acknowledgments

This paper is contribution number 235 from the Center for Water and the Environmental of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force.

References

[1] Balasubramanian, K. and Basak, S.C. (1998). Characterization of isospectral graphs using graph invariants and derived orthogonal parameters. *J. Chem. Inf. Comput. Sci.* **38,** 367–373.

[2] Balaban, A.T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89,** 399–404.

[3] Randić, M. (1975). On characterization of molecular branching. *J. Amer. Chem. Soc.* **97,** 6609–6615.

[4] Basak, S.C., Grunwald, G.D., and Niemi, G.J. (1997). Use of graph-theoretic and geometrical molecular descriptors in structure–activity relationships, In, *From Chemical Topology to Three-Dimensional Geometry* (A.T. Balaban, Ed.). Plenum Press, New York, pp. 73–116.

[5] Basak, S.C., Niemi, G.J., and Veith, G.D. (1990). Optimal characterization of structure for prediction of properties. *J. Math. Chem.* **4,** 185–205.

[6] Basak, S.C., Magnuson, V.R., Niemi, G.J., and Regal, R.R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **19,** 17–44.

[7] Basak, S.C., Bertelsen, S., and Grunwald, G.D. (1994). Application of graph theoretical parameters in quantifying molecular similarity and structure–activity relationships. *J. Chem. Inf. Comput. Sci.* **34,** 270–276.

[8] Basak, S.C. and Grunwald, G.D. (1999). Use of topological space and property space in selecting structural analogs. *Math. Modelling and Sci. Computing*, in press.

[9] Basak, S.C., Niemi, G.J., and Veith, G.D. (1990). Recent developments in the characterization of chemical structure using graph-theoretic indices. In, *Computational Chemical Graph Theory and Combinatorics* (D.H. Rouvray, Ed.). Nova, New York, pp. 235–277.

[10] Basak, S.C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med. Sci. Res.* **15,** 605–609.

[11] Basak, S.C. (1988). Binding of barbiturates to cytochrome P_{450}: A QSAR study using log *P* and topological indices. *Med. Sci. Res.* **16,** 281–282.

[12] Basak, S.C., Gute, B.D., and Grunwald, G.D. (1999). Development and application of molecular similarity methods: using nonempirical parameters. *Math. Modelling and Sci. Computing*, in press.

[13] Basak, S.C. and Gute, B.D. (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: A molecular similarity approach. In, *Proceedings of the 2nd International Congress on Hazardous Waste: Impact on Human and Ecological Health* (B.L. Johnson, C. Xintaras, and J.S. Andrews, Jr., Eds.). Princeton Scientific Publishing Co., Inc., New Jersey, pp. 492–504.

[14] Basak, S.C. and Grunwald, G.D. (1994). Molecular similarity and risk assessment: Analog selection and property estimation using graph invariants. *SAR QSAR Environ. Res.* **2,** 289–307.

[15] Auer, C.M., Nabholz, J.V., and Baetcke, K.P. (1990). Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure–activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **87,** 183–197.

[16] National Research Council (NRC). (1984). *Toxicity Testing: Strategies to Determine Needs and Priorities.* National Academy Press, Washington, D.C, p. 382.

[17] Arcos, J.C. (1987). Structure–activity relationships: Criteria for predicting carcinogenic activity of chemical compounds. *Environ. Sci. Technol.* **21,** 743–745.

[18] Trinajstić, N. (1992). *Chemical Graph Theory.* 2nd Edition. CRC Press, Boca Raton, Florida, p. 322.

[19] Kier, L.B. Murray, W.J., Randić, M., and Hall, L.H. (1976). Molecular connectivity V. connectivity series concept applied to density. *J. Pharm. Sci.* **65,** 1226–1230.

[20] Basak, S.C., Roy, A.B., and Ghosh, J. J. (1980). Study of the structure–function relationship of pharmacological and toxicological

agents using information theory. In, *Proceedings of the IInd International Conference on Mathematical Modelling* (Avula, X.J.R., Bellman, R., Luke, Y.L., and Rigler, A.K., Eds.). University of Missouri, Rolla, Missouri, Vol. II, pp. 851–856.

[21] Roy, A.B., Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications. In, *Mathematical Modelling in Science and Technology* (Avula, X.J.R., Kalman, R.E., Lipais, A.I., and Rodin, E.Y., Eds.). Pergamon Press. New York, pp. 745–750.

[22] Shannon, C.E (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**(2), 379–423.

[23] Wiener, N. (1948). *Cybernetics*. John Wiley and Sons, New York, p. 194.

[24] Ashby, W.R. (1956). *An Introduction to Cybernetics*. John Wiley and Sons, New York. p. 295.

[25] Kolmogorov, A.N. (1969). Combinatorial foundations of information theory and the calculus of probabilities. *Russian Math. Surveys.* **38**(2), 29–40.

[26] Rashevsky, N. (1955). Life, information theory and topology. *Bull. Math. Biophys.* **17,** 229–235.

[27] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **67,** 4517–4533.

[28] Bertz, S.H. (1981). The first general index of molecular complexity. *J. Am. Chem. Soc.* **103,** 3599–3601.

[29] Mowshowitz, A. (1968). Entropy and the complexity of graphs: I. and index of the relative complexity of a graph. *Bull. Math. Biophys.* **30,** 175–204.

[30] Trucco, E. (1956). On the information content of graphs: compound symbols; different states for each point. *Bull. Math. Biophys.* **18,** 237–253.

[31] Kier, L.B. (1980). Use of molecular negentropy to encode structure governing biological activity. *J. Pharm. Sci.* **69,** 807–810.

[32] Sarkar, R., Roy, A.B., and Sarkar P.K. (1978). Topological information content of genetic molecules—I. *Math. Biosci.* **39,** 299–312.

[33] Magnuson, V.R., Harriss, D.K., and Basak, S.C. (1983). Topological indices based on neighborhood symmetry: Chemical and biological applications. In, *Chemical Applications of Topology and Graph Theory* (R.B. King, Ed.). Elsevier, The Netherlands, pp. 178–191.

[34] Basak, S.C., Magnuson, V.R., Niemi, G.J., Regal, R.R., and Veith, B.D. (1987). Topological indices: Their nature, mutual relatedness, and applications. *Math. Model.* **8,** 300–305.

[35] Basak, S.C., Magnuson, V.R., Niemi, G.J., and Regal, R.R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **19,** 17–44.

[36] Basak, S.C., Frane, C.M., Rosen, M.E., and Magnuson, V.R. (1987). Molecular topology and mutagenicity: A QSAR study of monoketones. *Med. Sci. Res.* **18,** 887–888.

[37] Basak, S.C., Frane, C.M., Rosen, M.E., and Magnuson, V.R. (1986). Molecular topology and mutagenicity: A QSAR study of nitrosamines. *IRCS Med. Sci. Res.* **14,** 17–44.

[38] Basak, S.C. (1988). Binding of barbiturates to cytochrome P450: A QSAR study using log *P* and topological indices. *Med. Sci. Res.* **16,** 281–282.

[39] Hansch, C. (1976). On the structure of medicinal chemistry. *J. Med. Chem.* **19,** 1–6.

[40] Basak, S.C. Monsrud, L.J., Rosen, M.E., Frane, C.M., and Magnuson, V.R. (1986). A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharma Yugosld.* **36,** 81–95.

[41] Basak, S.C., Gute, B.D., and Ghatak, S. (1999). Prediction of complement inhibitory activity of benzamidines using topological and geometrical parameters. *J. Chem. Inf. Comput. Sci.* **39,** 255–260.

[42] Gute, B.D. and Basak, S.C. (1997). Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **7,** 117–131.

[43] Gute, B.C., Grunwald, G.D., and Basak, S.C. (1999). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHS): A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, in press.

[44] Basak, S.C., Gute, B.C., and Grunwald, G.D. (1997). Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. In, *Quantitative Structure–Activity Relationships in Environmental Sciences VII.* (R. Chen and G. Schuurmann, Eds.). SETAC Press, Pensacola, Florida, pp. 245–261.

APPENDIX 1.5 Normal boiling points of 1,ω-alkanedinitriles: The highest increment in a homologous series

Normal Boiling Points of 1,ω-Alkanedinitriles: The Highest Increment in a Homologous Series

Alexandru T. Balaban,*,[†] Subhash C. Basak,* and Denise Mills

Natural Resources Research Institute, University of Minnesota−Duluth, 5013 Miller Trunk Highway,
Duluth, Minnesota 55811

Received January 31, 1999

The normal boiling point for cyanogen is −22 °C; for its next homologue, malononitrile, it is 219 °C. The difference of 241 °C is apparently the highest one encountered for the addition of a single methylene group. Problems connected with boiling points and a rationalization for this observation are discussed in the context of intermolecular forces for liquids. A quantitative structure−property relationship (QSPR) study of the normal boiling points for monohaloalkanes and for the corresponding nitriles is reported. The behavior of the nitrile group as a pseudohalogen is also discussed. Normal boiling points of compounds having a cyano group bonded to an electron-attracting substituent situate the CN group close to being a pseudohalogen, but when the CN group is bonded to electron-donor substituents, the situation changes.

THE LIQUID STATE AND INTERMOLECULAR FORCES

Intermolecular forces range from the very weak ones such as those existing in liquefied noble gases to the strongest ones (hydrogen bonds) existing in hydrogen fluoride, in dimers of carboxylic acids (even in vapor state), or in liquids with multiple hydroxy groups such as glycols or water. The exceptional features of water (liquid state over a wide temperature range, expansion on freezing, high dielectric constant, and excellent solvent for a wide variety of substances) are responsible for making life possible on earth. Although ionic or metallic liquids also exist, they will not be discussed here because they are not molecular liquids. One should mention the important role of intermolecular forces and especially of hydrogen bonding in all life processes, in the transcription/translation processes involving DNA, in protein folding, receptor-agonist intercations, enzymatic mechanisms, etc.

Whereas intermolecular forces in crystals are compounded with conformational restrictions due to packing factors, liquids have molecular and conformational mobility (except for liquid crystals within certain limits). Liquids are more difficult to model than gases or solids. However, melting points of crystalline solids are also difficult to correlate with chemical structure due to packing factors, except for some classes of congeneric compounds.

Intermolecular forces are reflected by the following: vapor pressure versus temperature; boiling points at normal pressure (normal boiling points, NBPs); critical data; latent heat of vaporization versus temperature; viscosity; density and molar volume; optical properties such as the refractive index and molecular refractivity.

From all these clues, the easiest to measure with sufficient accuracy, and the most often cited for any compound, is the boiling point; usually, the NBP is cited, but seldom for compounds that would boil at temperatures above 250 °C at normal pressure because of decomposition. Many iodine derivatives decompose on heating even at lower temperatures because of the low C−I bond energy.

NITRILES AND THEIR NORMAL BOILING POINTS

The strongly electron-attracting nitrile (cyano) group is known to cause high dipole moments. For example, in the gas phase the dipole moments (in debye units) are as follows:[1]

for Me−X		for Ph−X	
X = Cl	1.87 D	X = Cl	1.70 D
X = CF_3	2.35 D	X = CF_3	2.86 D
X = NO_2	3.50 D	X = NO_2	4.21 D
X = CN	3.94 D	X = CN	4.39 D

The resulting dipole−dipole interactions lead to strong molecular associations, manifested in higher NBPs, heats of vaporization, and viscosities than those of the corresponding hydrocarbons with comparable molecular weights.

Among thermodynamic properties, normal boiling points have been extensively investigated in quantitative structure−property relationships (QSPRs). From the molecular descriptors used in such correlations, topological indices have been among the most successful.[2−6] For alkanes, such QSPR studies allow nowadays the prediction of NBPs within a range of 2 or 3 °C.[7−9] For various other classes of compounds many QSPR studies are available, and their accuracy range is often lower than 10 °C.[10−15]

Nitriles, however, proved to defy simple approaches. Thus, a recent study by Wessel and Jurs for a diverse set of industrially important chemicals containing nitrogen with mean-square-root errors of about 9 °C led to satisfactory results for mononitriles but to very large errors for two dinitriles, namely, cyanogen and malononitrile.[12] We have therefore decided to look more closely into this matter. A comprehensive review on malononitrile is available.[16]

[†] Permanent address: Department of Organic Chemistry, Polytechnic University Bucharest, Romania.

10.1021/ci9900074 CCC: $18.00 © 1999 American Chemical Society
Published on Web 09/04/1999

Table 1. Cyano Group as a Pseudohalogen: NBPs for X–Y or X_2 Compounds[a]

(pseudo)-halogen X	FW	X–CN NBP (°C)	X–CN FW	X–X NBP (°C)	X–X FW
F	19	−72	49	−188	38
CN	26	−22	60	−22	60
Cl	35	13	66	−35	71
Br	80	62	110	56	160
I	127	184	157	178	254

[a] Figures have been rounded off to the nearest integer.

Table 2. NBPs of Cyanotrihalomethanes $Hal_3C–CN$ and of Tetrahalomethanes $Hal_3C–X$ (Hal = F, Cl, Br)[a]

X	F NBP (°C)	F FW	Cl NBP (°C)	Cl FW	Br NBP (°C)	Br FW
F	−128	88	25	137	107	271
Cl	−82	104	77	154	160	287
CN	−62	95	84	149	170	278
Br	−79	149	104	198	190	332
I	−23	196	141	245		

[a] Figures have been rounded off to the nearest integer.

CYANO GROUP AS A PSEUDOHALOGEN

Groups such as cyano, thiocyano, cyanato, and azido are considered to be pseudohalogens.[17−19] In this paper we shall focus only on the cyano group. There are also significant differences, however, between some compounds of halogens and pseudohalogens, for instance the fact that hydrogen cyanide is a much weaker acid (with $pK_a = 9.2$) than hydrogen halides. Also, the coordinating ability of the cyanide anion for iron leads to a high toxicity, whereas each of the halide anions has a different biological significance. One should also recall that the cyano group is bidentate, being able to form covalent or coordinative bonds at the carbon or nitrogen atoms. Thus, the elongated shape of the cyano group makes it different from the spherical halogens.

It is known that molecular weights have a large influence on NBPs. According to its formula weight (FW), a CN group is intermediate between a fluorine and a chlorine atom. On comparing NBPs[20−22] of simple halogens, interhalogens, cyanogen, or cyanogen halide linear molecules (Table 1), it can be seen that the cyano group does indeed behave as a pseudohalogen. On considering cyanogen halides, the CN group is placed by NBPs between fluorine and chlorine. However, on comparing NBPs of cyanogen and those of elemental halogens, the CN group is situated between chlorine and bromine, as if the CN group had a slightly higher formula weight.

In Table 2 the NBPs of cyanotrihalomethanes, $X_3C–CN$, and of tetrahalomethanes, CX_4, are shown. It can be seen that the cyano group behaves again as a pseudohalogen situated between chlorine and bromine.

Although some physical data support the idea that the CN group manifests itself as a pseudohalogen, its chemical behavior in organic compounds is quite different from that of halogens. The C–Cl, C–Br, and C–I bond strengths are much lower than the bond strength of the C–CN bond; therefore, these halogens (unlike CN groups) are good leaving groups. In the next section we shall examine organic compounds whose NBPs are much higher than those of the corresponding halogen compounds, so that the cyano group would be situated beyond iodine; in such cases, the notion of pseudohalogen is no longer justified.

NORMAL BOILING POINTS OF NITRILES AND DINITRILES

Mononitriles have NBPs which are quite high when compared with the corresponding halides (Table 3). In Table 3 structures of halogen derivatives are indicated (in abbreviated form) according to IUPAC nomenclature rules; for nitriles, however, to achieve consistency, the CN group is considered as a pseudohalogen; therefore, the nomenclature is no longer according to IUPAC. In these cases a CN group increases the NBP much more than the heaviest stable halogen atom, namely, iodine. An analogous behavior is apparent when comparing halocarbonyl or cyanocarbonyl compounds (Table 4). Also, the NBPs of 1,*ω*-alkanedihalides for linear alkane chains with one through four carbon atoms, $X(CH_2)_nX$ (with $n = 1-4$) are much lower than for the corresponding 1,*ω*-alkanedinitriles (Table 5).

As seen from Table 6 for *gem*-dihalides or *gem*-dinitriles of methane, ethane, or propane, a similar trend with higher NBPs for X = CN than for X = Hal is observed; moreover, one sees the curious trend that when the X group in R–X is I or CN, the NBPs decrease progressively in the above series with increasing molecular weight, whereas the corresponding compounds with X = F, Cl, or Br exhibit the reverse, normal behavior. A break in Table 6 separates the compounds with normal and abnormal behavior.

QSPR STUDY OF MONOHALO DERIVATIVES AND OF THEIR CYANO ANALOGUES

For correlating the chemical structure with the NBP for the data presented in Table 3 we selected eleven topological indices: the information indices $IC_1–IC_3$ and $CIC_1–CIC_3$;[23] the Wiener index *W*; the valence connectivity indices $^0\chi^v$–$^2\chi^v$;[4,24] and the average distance-sum connectivity adapted for heteroatoms based on their electronegativities (Balaban's index, J_x).[25,26] All indices except the last one were computed using the program POLLY.[27]

Due to the fact that the scale of the various topological indices may differ by several orders of magnitude, all indices were transformed by first adding 1 to the index and then taking the natural logarithm of this result. The transformed version of the indices was used in all analyses. The CORR procedure of the SAS statistical package[32] was used to identify intercorrelated indices. The elimination of such indices reduced to four the number of selected TIs, namely, IC_2, CIC_2, $^1\chi^v$, and J_x.

An all-subset regression was accomplished using the REG procedure of the same statistical package,[32] which indicated that $^1\chi^v$ and J_x gave the best results; IC_2 and CIC_2 gave the next best results. The drawback of IC and CIC indices is that the nature of the halogen does not affect the value of these indices.

Experimental and calculated data for NBPs of monohalo derivatives with one through five carbon atoms and the corresponding mononitriles with two to six carbon atoms are presented in Table 3, above the solid line. Some nitriles with six to eight carbon atoms are also included below the solid line, but they have no halogen counterparts, and the correlations discussed below do not include them.

Table 3. NBPs (°C) of Organic Halides and Nitriles R−X and QSAR in Terms of $^1\chi^v$ and J_x

compd	NBP$_{exp}$[a]	NBP$_{calc}$[a]	diff$_{expt-calc}$[a]	$^1\chi^{v}$[b]	J_x[b]
Me−F	−78	−81	3	0.3206	0.6054
Et−F	−38	−32	−6	0.6801	0.9143
Pr−F	3	3	0	0.9058	1.0550
Bu−F	33	34	−1	1.0899	1.1346
sBu−F	25	22	3	1.0685	1.2334
1-C$_5$−F	63	62	1	1.2453	1.1860
Me−Cl	−24	−23	−1	0.7580	0.6152
Et−Cl	12	10	2	0.9199	0.9207
Pr−Cl	47	47	0	1.1016	1.0588
iPr−Cl	36	34	2	1.0328	1.1656
Bu−Cl	79	79	0	1.2553	1.1375
sBu−Cl	68	69	−1	1.2081	1.2369
iBu−Cl	69	70	−1	1.2134	1.2407
tBu−Cl	51	51	0	1.1207	1.3635
1-Cl−C$_5$	108	106	2	1.3885	1.1881
2-Cl−C$_5$	97	97	0	1.3473	1.2672
2-Me−1-Cl−C$_4$	100	100	0	1.3617	1.3043
3-Me−1-Cl−C$_4$	99	98	1	1.3520	1.2709
CEt$_2$−Cl	98	99	−1	1.3571	1.2999
Me−Br	4	10	−6	1.0865	0.6403
Et−Br	39	33	6	1.1301	0.9357
Pr−Br	71	69	2	1.2798	1.0685
iPr−Br	60	56	4	1.1906	1.1768
Bu−Br	102	99	3	1.4100	1.1445
sBu−Br	91	90	1	1.3421	1.2453
iBu−Br	91	97	−6	1.3742	1.2479
tBu−Br	73	77	−4	1.2476	1.3727
1-Br−C$_5$	130	125	5	1.5252	1.1936
2-Br−C$_5$	117	116	1	1.4649	1.2737
2-Me−Br−1-C$_4$	121	125	−4	1.5019	1.3100
3-Me−1-Br−C$_4$	120	122	−2	1.4934	1.2765
CEt$_2$−Br	119	119	0	1.4736	1.3070
Me−I	43	51	−8	1.2627	0.6689
Et−I	73	66	7	1.2528	0.9532
Pr−I	103	100	3	1.3863	1.0801
iPr−I	90	87	3	1.2862	1.1900
Bu−I	131	127	4	1.5041	1.1531
sBu−I	120	117	3	1.4248	1.2553
iBu−I	121	127	−6	1.4716	1.2568
tBu−I	100	106	−6	1.3265	1.3833
1-I−C$_5$	155	150	5	1.6094	1.2000
2-I−C$_5$	141	141	0	1.5384	1.2818
2-Me−1-I−C$_4$	148	153	−5	1.5880	1.3169
3-Me−1-I−C$_4$	147	149	−2	1.5802	1.2829
CEt$_2$−I	146	145	1	1.5465	1.3156
Me−CN	82	71	11	0.5446	1.2196
Et−CN	97	104	−7	0.8259	1.2173
Pr−CN	118	126	−8	1.0239	1.2366
iPr−CN	104	109	−5	0.9810	1.3592
Bu−CN	141	143	−2	1.1891	1.2565
sBu−CN	125	128	−3	1.1647	1.3880
iBu−CN	131	129	2	1.1442	1.3483
tBu−CN	106	108	−2	1.0899	1.5065
1-CN−C$_5$	164	158	6	1.3308	1.2737
2-CN−C$_5$	146	145	1	1.3097	1.3888
2-Me−1-CN−C$_4$	154	147	7	1.2920	1.3431
3-Me−1-CN−C$_4$	157	158	−1	1.3308	1.2737
CEt$_2$−CN	146	142	4	1.3199	1.4339
EtCMe$_2$−CN	129	126	3	1.2624	1.5304
1-CN−C$_6$	183	171	12	1.4549	1.2881
2-CN−C$_6$	164	160	4	1.4363	1.3840
3-Me−1-CN−C$_5$	172	158	14	1.4298	1.3992
4-Me−1-CN−C$_5$	180	159	21	1.4298	1.3830
5-Me−1-CN−C$_5$	180	158	22	1.3308	1.2737
1-CN−C$_7$	199	183	16	1.5653	1.3002

[a] Figures have been rounded off to the nearest integer. [b] Topological indices $^1\chi^v$ and J_x are expressed by converting their values (y) into $\ln(1 + y)$.

Table 4. NBPs of Halocarbonyl Derivatives (Iodine Derivatives Are Not Available)[a]

	NBP (°C)	
X	EtOCOX	ClCOX
F	57	−45
Cl	95	8
Br	116	25
CN	116	128

[a] Figures have been rounded off to the nearest integer.

Table 5. NBPs of 1,ω-Dihalides and 1,ω-Biscyanides of Linear Alkanes C$_1$−C$_4$[a]

	NBP (°C)			
X	XCH$_2$X	X(CH$_2$)$_2$X	X(CH$_2$)$_3$X	X(CH$_2$)$_4$X
F	−52	31	42	78
Cl	40	84	121	154
Br	97	131	167	197
I	181	200	227	
CN	219	266	286	295

[a] Figures have been rounded off to the nearest integer.

Table 6. NBPs of *gem*-Bis(pseudo)halides of Alkanes C$_1$−C$_3$[a]

	NBP (°C)		
X	CH$_2$X$_2$	MeCHX$_2$	Me$_2$CX$_2$
F	−52	−25	0
Cl	40	58	71
Br	97	113	115
I	181	178	148
CN	219	198	170

[a] Figures have been rounded off to the nearest integer.

A comment on how the $^1\chi^v$ and J_x indices vary with increasing size and branching of molecules needs to be added. Both these indices increase with increasing size. The nature of the halogen X in R−X molecules with the same R group also leads to a progressive increase in the series F, Cl, Br, and I; this increase is steep for $^1\chi^v$ but moderate for J_x. However, increasing branching of the R group for isomeric molecules leads to decreasing values for $^1\chi^v$ but to increasing values for J_x. Of course, as a general rule, experimental NBPs increase with increasing size and molecular weight of molecules and decrease with molecular branching; only poly(fluoroalkanes) are exceptions to this rule, as mentioned earlier.[13]

The corresponding equations are shown in Table 7a,b with the statistical parameters. For the chloro derivatives J_x was not a significant parameter, so that a monoparametric equation in terms of $^1\chi^v$ gave in this case satisfactory results. For all other compounds from Table 3, such monoparametric equations led to worse results than those presented in both parts a and b of Table 7. Intercorrelation factors between the four selected indices are presented in Table 8; one can see that no significant intercorrelation is present. It can be observed from Tables 3 and 7a that the correlation for nitriles is slightly poorer than for the halogens; however, the agreement between the experimental and calculated NBPs is quite good. Remarkably, the coefficients of the $^1\chi^v$ parameter are similar for Br and I in Table 7a and for all halogens in Table 7b; this fact is reminiscent of the observation presented in the earlier paper[13] about the fact

Table 7. Correlation Equations for NBP and Statistical Parameters

(a) In Terms of $^1\chi^v$ and J_x

	NBP	s	r	F
RF	$(208 \pm 23)^1\chi^v - (84.0 \pm 32)J_x - (96.8 \pm 15)$	4.3	0.998	355
RCl	$(204 \pm 2)^1\chi^v - (177 \pm 2)$	1.4	0.999	9444
RBr	$(203 \pm 12)^1\chi^v + (45.6 \pm 9.3)J_x - (239 \pm 12)$	4.5	0.994	404
RI	$(195 \pm 15)^1\chi^v + (59.3 \pm 10)J_x - (235 \pm 17)$	5.3	0.989	235
RCN	$(117 \pm 8.4)^1\chi^v - (92.5 \pm 22)J_x + (120 \pm 27)$	6.1	0.976	99

(b) In Terms of IC_2 and CIC_2

	NBP	s	r	F
RF	$(223 \pm 27)IC_2 + (197 \pm 47)CIC_2 - (406 \pm 38)$	10.2	0.988	62
RCl	$(230 \pm 16)IC_2 + (146 \pm 16)CIC_2 - (333 \pm 27)$	9.0	0.978	109
RBr	$(218 \pm 16)IC_2 + (135 \pm 16)CIC_2 - (286 \pm 27)$	8.7	0.977	104
RI	$(198 \pm 15)IC_2 + (115 \pm 15)CIC_2 - (217 \pm 25)$	8.4	0.974	92
RCN	$(176 \pm 30)IC_2 + (93.9 \pm 20)CIC_2 - (168 \pm 42)$	11.4	0.914	25

Table 8. Intercorrelation Matrix for the Four Selected TIs[a]

	$^1\chi^v$	J_x	IC_2	CIC_2
$^1\chi^v$	1.000	0.702	0.802	0.178
J_x		1.000	0.451	0.655
IC_2			1.000	−0.331
CIC_2				1.000

[a] Topological indices $^1\chi^v$ and J_x are shown by converting their values (y) into $\ln(1 + y)$.

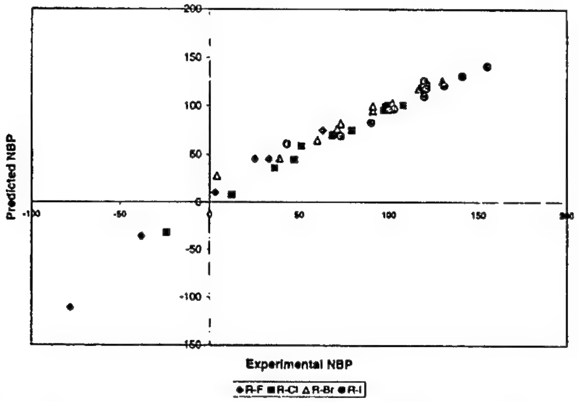


Figure 1. Plot of the predicted NBP versus the experimental NBP for the combined set of 45 monohaloalkanes from Table 3 in terms of two TIs, namely, $^1\chi^v$ and J_x.

that one might consider a "generalized halogen" with a stepwise increment for the four halogens F, Cl, Br, and I. Though the aim of the present paper was to discuss nitriles and not haloderivatives (the NBPs of these last compounds were the object of a QSPR study in the earlier paper[13]), one can use the same parameters as in Table 7a for a correlation of NBPs for all 45 halogen derivatives presented in Table 3 according to the following equation:

$$NBP = (180 \pm 7.8)^1\chi^v + (34 \pm 10)J_x - (189 \pm 9.2)$$
$$s = 10\ °C \qquad r = 0.9823 \qquad F = 579$$

The diagram shown for this correlation in Figure 1 indicates that only 2-butyl fluoride and three halomethanes with F, Br, and I have deviations above 14 °C between observed and predicted NBPs.

Interestingly, the last equation of Table 7a works even for other aliphatic mononitriles with six to eight carbon atoms, presented at the bottom of Table 3 below the full line; however, in these cases, all calculated values are lower than the experimental ones.

Table 9. NBPs (°C) of Unsaturated Nitriles and QSAR in Terms of IC_2 and CIC_2

nitrile	IC_2	CIC_2	NBP_{exp}	NBP_{calc}	$diff_{expt-calcd}$
C=CC#N	1.2590	0.2515	78	80	−2
C#CC#N	1.2006	0.0000	43	40	3
C=CCC#N	1.3666	0.3365	119	112	7
CC=CC#N	1.2936	0.5158	113	116	−3
CC(=C)C#N	1.2936	0.5158	91	116	−25
CC=C(C)C#N	1.2548	0.7853	138	137	1
CC(C)=CC#N	1.2102	0.8531	141	135	6
C=CCCC#N	1.3689	0.5704	140	138	2
CCC=CC#N	1.3930	0.5146	136	137	−1
C=CC=CC#N	1.3468	0.4787	137	123	14
CCC(C)=CC#N	1.3625	0.7391	142	155	−13
CC(C)C=CC#N	1.3300	0.7971	155	155	0
CC(C)=CCC#N	1.3300	0.7971	166	155	11

Unsaturation in the nitrile chain lowers appreciably the NBP, as seen in Table 8. Using the same descriptors as in Table 7b for these nitriles with three to six carbon atoms having one or two double bonds or one triple bond (denoted by # in Table 9 which uses Smiles notation for structures), the QSAR results presented in Table 9 were obtained with the following equation:

$$NBP = (214 \pm 52)IC_2 + (109 \pm 13)CIC_2 - (217 \pm 67)$$
$$s = 11\ °C \qquad r = 0.9121 \qquad F = 52$$

A GUESSING GAME

On addressing an audience of chemists, the following guessing game was proposed: the audience was given the NBPs of the 1,ω-alkanedinitriles $X(CH_2)_nX$ with $n = 1-4$, namely, malononitrile, succinonitrile, adiponitrile, and capronoonitrile (i.e., the last line in Table 5). Then everyone was asked to guess the NBP temperature interval for oxalonitrile (the compound with $n = 0$) by putting a mark in one of the following eight intervals: <−20; −20 to +20; +20 to +60; +60 to +100; +100 to +140; +140 to +180; +180 to +220; and >+220 °C. Remarkably, no member of the audience guessed that oxalonitrile (cyanogen with NBP = −22 °C) should appear in the first temperature interval (NBP ≤ −20 °C). The other seven temperature intervals were about equally populated with marks.

LARGEST INCREMENT IN NBP FOR A HOMOLOGOUS SERIES

The two compounds (cyanogen and manononitrile) mentioned to be outliers in the QSPR study cited earlier[12] represent the pair with the largest NBP increment on adding one methylene group, as seen from Table 10. In this table, one compares the next two homologues having various simple groups bonded either directly (R_2) or via a methylene group (RCH_2R), where R can be a halogen, a cyano group, an alkyl, an alkoxy, or an organic electronegative group. Breaks in the table delineate various related classes of compounds.

The first entry of the above two compounds constitutes a class by itself. The huge difference of 241 °C between the NBPs of cyanogen (oxalonitrile, with NBP = −22 °C) and malononitrile (with NBP = 219 °C) can be explained by the fact that cyanogen has a linear geometry and hence a

Table 10. Differences in NBPs for Compounds Differing by One Methylene Group[a]

	NBP (°C)		
R	R_2	RCH_2R	diff
CN	−22	219	241
H	−253	−162	91
CF_3	−78	1	79
CH_3-CO	88	138	50
CH_3	−89	−42	47
HC≡C	10	55	45
F	−188	−52	136
Cl	−35	40	75
Br	56	97	41
I	184	182	−2
Et	0	37	37
MeO	14	42	28
EtO	63	88	25
MeS	110	149	39
EtS	154	181	27
CCl_3	186	206	20
COOMe	163	181	18
COOEt	185	199	14
COOPr	211	229	18
COOBu	242	256	14
Ph	256	264	12

[a] Figures have been rounded off to the nearest integer.

zero dipole moment, whereas malononitrile is a V-shaped molecule with a high dipole moment, 3.58 D.[28,29] The calculated polarizability of malononitrile is abnormally high in comparison with calculated values.[30,31]

A few other comments in Table 10 should be added. The first nine entries show differences in NBPs that are higher than 40 °C for the two homologues. Among these, the first six have electronegative or slightly electron-donating groups; the next class includes the four stable halogens, and the trend in this group with progressively decreasing electronegativity is quite interesting, starting with the next highest NBP difference in the whole table (for fluorine) and ending with a negative difference (for iodine). All these entries have linear R_2 and bent R_2CH_2 molecules for the two homologues, respectively.

The last class with NBP differences lower than 40 °C, however, demonstrates that electronegativity by itself does not provide a full explanation for the data contained in Table 10. Indeed, here again we encounter groups with electron-donating as well as with electron-accepting properties. However, in this class the R_2 molecules have no longer linear geometries except for biphenyl and hexachloroethane.

OTHER DINITRILES

A comparison between volatilities of dinitriles of four-carbon dicarboxylic acids is interesting, despite the incompletely matched data. Succinonitrile has a NBP of 266 °C and a dipole moment of 3.93 D. From the two stereoisomeric olefinic congeners, the dinitrile of fumaric acid with *E*-geometry is more volatile (NBP of 186 °C, subliming even under 100 °C) than the dinitrile of maleic acid (with a higher dipole moment because of its Z-geometry) which has a BP of 111 °C *at 20 Torr* and 99 °C *at 13 Torr*. The alkynic congener which has a linear geometry and zero dipole moment (dicyanoacetylene or acetylenedicarbonitrile, C_4N_2) has a NBP of only 77 °C and sublimes easily. Interestingly, the dinitrile C_6N_2 of hexadiynedioic acid with two triple bonds (with linear geometry) has a NBP of only 154 °C.

Isomers of benzodinitrile also have volatilities that attest the importance of dipole moments: phthalonitrile with the highest dipole moment has *at 10 Torr* a boiling point of 151 °C; isophthalonitrile with a dipole moment which is about half as large has the BP of 140 °C at the same reduced pressure; and terephthalonitrile with a zero dipole moment sublimes *at normal pressure* at temperatures starting at 153 °C.

When the CN group is attached to an electron-acceptor substituent, the polarity of the bond is low and the NBP is within the range expected for a pseudohalogen with a formula weight close to that of chlorine. However, when the CN group is bonded to an electron-donor substituent, the high polarity of the resulting bond enhances appreciably the NBP. The conclusion is that NBPs are the result of a multiplicity of factors inherent in determining the intermolecular forces that exist in the liquid state. In certain cases such as the two homologous dinitriles with two and three carbon atoms, QSPR studies should not ignore differences between these intermolecular interactions.

ACKNOWLEDGMENT

This is contribution number 250 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force. The assistance of G. D. Grunwald and B. D. Gute is acknowledged.

REFERENCES AND NOTES

(1) Grundnes, J.; Klaboe, P. Basicity, Hydrogen Bonding and Complex Formation. In *The Chemistry of the Cyano Group*; Rappoport, Z., Ed.; Interscience-Wiley: London, 1970; pp 123−166.

(2) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure−activity correlations. *Top. Curr. Chem.* **1983**, *114*, 21−55.

(3) Trinajstic, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992; pp 225−274.

(4) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.

(5) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press-Wiley: Chichester, U.K., 1993.

(6) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.

(7) Balaban, A. T. Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.

(8) Devillers, J., Balaban, A. T., Eds. *Topological Indices and Related Molecular Descriptors for QSAR and QSPR Studies*; Gordon & Breach: New York, in press.

(9) Balaban, A. T.; Balaban, T. S. Correlations using topological indices based on real graph invariants. *J. Chim. Phys. Phys.-Chim. Biol.* **1992**, *89*, 1735−1745.

(10) Stanton, D. T.; Jurs, P. C. Computer assisted prediction of normal boiling points of furans, tetrahydrofurans, and thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301−310.

(11) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 639−645.

(12) Wessel, M. D.; Jurs, P. C. Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.

(13) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations between chemical structure and normal boiling points of halogenated alkanes C_1-C_4. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233−237.

(14) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between chemical structure and normal boiling points of acyclic ethers, peroxides, acetals, and their sulfur analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237−244.

(15) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between structure and normal boiling points of haloalkanes C_1-C_4 using neural networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118−1121.

(16) Freeman, F. The chemistry of malononitrile. *Chem. Rev.* **1969**, *69*, 591−624.

(17) King, R. B., Ed. *Encyclopedia of Inorganic Chemistry*, Vol. 2; Wiley: Chichester, 1994; p 558.

(18) Gutmann, V., Ed. *Halogen Chemistry*; Academic Press: New York, 1967.

(19) Haas, A. The element displacement principle, a new guide in p-block element chemistry. *Adv. Inorg. Chem.* **1984**, *28*, 167−202. Haas, A.; Brosius, A. Influence of fluorine and parahalogen substituents on the chemistry of some functional groups. *ACS Symp. Ser.* **1994**, *555*, 104−127.

(20) Lide, D. R. *Handbook of Chemistry and Physics*, 79th ed., 1998−1999; CRC Press: Boca Raton, FL, 1988.

(21) Weast, R. C.; Astle, M. J. *Handbook of Data on Organic Compounds*; CRC Press: Boca Raton, FL, 1985.

(22) Daubert, T. E., Danner, R. P., Eds. Design Institute for Physical Property Data (DIPR). *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*; Hemisphere: New York, 1989; Vols. 1−4.

(23) Basak, S. C.; Magnuson, V. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.

(24) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(25) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *80*, 399−404.

(26) Balaban, A. T. Chemical graphs. 48. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.

(27) Basak, S. C.; Harris, D. K.; Magnuson, V. R. *POLLY 2.3*; University of Minnesota: Duluth, MN, 1988.

(28) van der Kelen, G. P. Dipole moments and infrared spectra of haloacetonitriles. *Bull. Soc. Chim. Belg.* **1962**, *71*, 421 (*Chem. Abstr.* **1964**, *58*, 2934).

(29) Schwarz, M.; Kitchman; L. A.; Tucker, R. W.; Nelson, E. R. Dielectric constants, dipole moments and structures of dinitriles. *J. Chem. Eng. Data* **1970**, *15*, 341 (*Chem. Abstr.* **1970**, *72*, 131866).

(30) Le Fèvre, R. J. W.; Orr, B. J.; Ritchie, G. L. D. Molecular polarisability. Anisotropic polarisability of the cyano-group from molar Kerr constants and dipole moments of eight nitriles. *J. Chem. Soc.* **1965**, 2499−2505.

(31) Lippincott, E. R.; Nagarajan, G.; Stutman, J. M. Polarizabilities from the δ-function model of chemical binding. II. Molecules with polar bonds. *J. Chem. Phys.* **1966**, *70*, 78−84.

(32) *SAS/STAT User's Guide*, Release 6.03 ed.; SAS Institute, Inc.: Cary, NC, 1988.

CI9900074

APPENDIX 1.6 A comparative QSAR study of benzamidines complement-inhibitory activity and benzene...

ELSEVIER

Computers and Chemistry 24 (2000) 181–191

Computers & Chemistry

www.elsevier.com/locate/compchem

A comparative QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity

Subhash C. Basak [a,*], Brian D. Gute [a], Bono Lučić [b], Sonja Nikolić [a,b], Nenad Trinajstić [a,b]

[a] *Natural Resources Research Institute, The University of Minnesota, Duluth, MN 55811, USA*
[b] *The Rugjer Bošković Institute, PO Box 1016, HR-10001 Zagreb, Croatia*

Received 5 July 1999; accepted 20 September 1999

Abstract

A novel QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity is reported and a new efficient method for selecting descriptors is used. Complement–inhibitory activity QSAR models of benzamidines contain from one to five descriptors. The best, according to fitted and cross-validated statistical parameters, is shown to be the five-descriptor model. Models with a higher number of indices did not improve over the five-descriptor model. The benzene derivatives structure–toxicity models involve up to seven linear descriptors. Multiregression models, containing up to ten nonlinear descriptors, are also reported for the sake of comparison with previously obtained additivity models. Comparison with benzamidine complement–inhibitory activity models and with benzene derivatives toxicity models from the literature favors our novel approach. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: QSAR study; Complement–inhibitory activity; Benzene; Five-descriptor model

1. Introduction

In our recent papers a hierarchical QSAR (quantitative structure–activity relationship) approach was used to model the complement–inhibitory activity of benzamidines (Basak et al., 1999a) and the acute aquatic toxicities of benzene derivatives (Gute and Basak, 1997; Basak et al., 1999c). The hierarchical QSAR approach uses topological (partitioned into topostructural and topochemical), geometric and quantum-chemical descriptors in a stepwise fashion to build increasingly more complex structure–property–activity models (Basak et al., 1997, 1999b). Now we report the use, with the same aim, of a new efficient approach for selecting the best QSAR models using multivariate regression (MR) (Lučić and Trinajstić, 1999; Lučić et al., 1999a) and a standard approach for variable selection and model generation used in CODESSA (Katritzky et al., 1999; Lučić et al., 1999b). Sometime ago Hansch and Yoshimoto (Hansch and Yoshimoto, 1974) carried out a QSAR study on the complement–inhibitory potency of benzamidines using their own approach. After 10 years, Hall et al. (Hall et al., 1984) carried out a QSAR study on the toxicities of benzene derivatives using de novo analysis (Free and Wilson, 1964; Kubinyi and Kehrhahn, 1976), and derived an additivity model for 66 compounds (they excluded three compounds as outliers). We will analyze their models and compare to ours.

[*] Corresponding author.

PII: S0097-8485(99)00059-5

Benzamidines are inhibitors of the complement system. Complement is a system of factors occurring in normal serum which are characteristically activated by antibody–antigen interactions and which subsequently mediate a number of biologically significant consequences. The factors of the complement system include at least 20 chemically distinct serum proteins and glycoproteins. These factors which normally exist in an inactive form, may be activated by two (classical and alternative) pathways. Both pathways generate macromolecular membrane attack complexes which lyse a variety of cells, bacteria and viruses (Kuby, 1992). Products of this activation result in inflammatory reactions at the site of antibody–antigen interaction. This is especially pronounced in the case of organ specific and systemic autoimmune disorders. Therefore. control of unregulated complement activation is essential, especially in the case of autoimmune disease.

Acute aquatic toxicities of benzene derivatives in the fathead minnow (*Pimephales promelas*) indicate 96-h values ranging from 3.0 to 6.4 log units for the LC50 (lethal dose to 50% of the sample). Details about LC50 measurements are given in the report by Hall et al. (Hall et al., 1984).

2. Data sets

2.1. Benzamidines

In Fig. 1 we give the structural formula of benzamidines and in Table 1 the side-chain structures and experimental complement–inhibitory activities in terms of 1/log *C* for studied benzamidines. *C* in log *C* is the micromolar concentration of inhibitor required for 50% inhibition of lyophilized guinea pig complement when assayed in buffer (Hansch and Yoshimoto. 1974).

2.2. Benzene derivatives

Toxicity data of 69 benzene derivatives are taken from Hall et al. (Hall et al., 1984). Toxicity data reported by Hall et al. consists of 26 original experimental observations and 43 taken from seven different sources. Thus, the studied set of benzene derivatives contains toxicities of 68 compounds and benzene. The benzene derivatives in this set have seven different substituents; each substituent being present in at least six compounds. These substituents are amino, bromo, chloro, hydroxyl, methyl, methoxyl and nitro groups. Studied benzene derivatives are listed in Table 2. Their toxicities are expressed as the negative logarithm of the lethal concentration of a benzene derivative and denoted by $-\log(LC_{50})$.

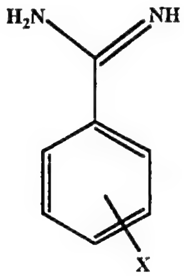


Fig. 1. Structural formula of benzamidines.

2.3. Molecular descriptors

In Table 3 are given symbols and brief description of descriptors that are used for the QSAR modeling of benzamidines and benzene derivatives in the present work. The total number of descriptors is 110 (40 topostructural, 61 topochemical, three geometric and six quantum-chemical descriptors). In the previous QSAR study of benzamidines (Basak et al., 1999a) 95 descriptors were used (37 topostructural, 55 topochemical and three geometric). The difference is caused by a fact that nine topological descriptors possess zero values (we included them in our set simply to have the complete set of descriptors) for all molecules studied and six quantum-chemical descriptors were not included in the previous modeling. All topological descriptors were transformed as it was done Basak et al. (Basak et al., 1999a) using a natural logarithmic transformation of the form $\ln(x+1)$, where x represents single values of descriptors. This was done to avoid errors in rounding up numerical values because the range of descriptor values was rather large. The geometric descriptors were transformed by the natural logarithm of the descriptor for consistency.

In the case of benzene derivatives we used the same set of descriptors as Gute and Basak (Gute and Basak, 1997) and Basak et al. (Basak et al., 1999c). They were transformed in the same way as the benzamidine data set (see Basak et al., 1999a).

2.4. Variable selection and models generation

To obtain the best possible QSAR models with I ($I = 1, 2, 3, \ldots$) descriptors we used a computational approach, detailed elsewhere (Lučić and Trinajstić, 1999), by which one can select the best MR model with I descriptors from the set of N descriptors. The number of possible models with I descriptors is $N!/(N-I)!I!$ The quality of each model (with I descriptors) was identified with its correlation coefficient (R), and among all possible models the best one was selected, with the highest value of R. To be able to check the quality of a large number of MR models, it was necessary to develop a very fast procedure for calculating R,

Table 1
Observed and calculated (cross-validated, CV, and fitted, FIT) complement-inhibitory activities 1/log C of 105 benzamidines

No.	X	1/log C		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
1	2-CH_3	−0.444	−0.417	−0.419
2	3,4-$(CH_3)_2$	−0.425	−0.423	−0.424
3	H	−0.418	−0.424	−0.423
4	3-OH	−0.415	−0.439	−0.434
5	3-CF_3	−0.410	−0.378	−0.382
6	3-NO_2	−0.410	−0.392	−0.395
7	3-Br	−0.405	−0.399	−0.400
8	3-CH_3	−0.398	−0.399	−0.399
9	3-OCH_3	−0.397	−0.401	−0.401
10	3-$CH_2C_6H_5$	−0.373	−0.343	−0.346
11	3,5-$(CH_3)_2$	−0.361	−0.375	−0.369
12	3-OC_3H_7	−0.355	−0.358	−0.358
13	3-i-C_5H_{11}	−0.355	−0.344	−0.345
14	3-OC_4H_9	−0.351	−0.340	−0.341
15	3-C_4H_9	−0.338	−0.355	−0.353
16	3-$CH{=}CHC_6H_5$	−0.339	−0.324	−0.325
17	3-$OCH_2C_6H_5$	−0.331	−0.324	−0.324
18	3-$(CH_2)_2C_6H_5$	−0.330	−0.332	−0.331
19	3-OC_6H_{13}	−0.329	−0.318	−0.319
20	3-$O(CH_2)_4OC_6H_5$	−0.325	−0.286	−0.287
21	3-$O(CH_2)_2OC_6H_5$	−0.323	−0.314	−0.315
22	3-C_6H_5	−0.323	−0.366	−0.359
23	3-$O(CH_2)_3OC_6H_4$-4-COOH	−0.321	−0.296	−0.297
24	3-OC_5H_{11}	−0.320	−0.327	−0.326
25	3-O-i-C_5H_{11}	−0.318	−0.338	−0.335
26	3-$O(CH_2)_2OC_{10}H_7$-α	−0.312	−0.255	−0.262
27	3-$O(CH_2)_4OC_6H_4$-4-NH_2	−0.306	−0.288	−0.289
28	3-$(CH_2)_4C_6H_5$	−0.302	−0.315	−0.313
29	3-$O(CH_2)_3OC_6H_4$-4-NO_2	−0.301	−0.282	−0.282
30	3-$O(CH_2)_3OC_6H_4$-4-NH_2	−0.300	−0.298	−0.298
31	3-$(CH_2)_2$-4-C_5H_4N	−0.299	−0.318	−0.318
32	3-$O(CH_2)_3OC_6H_5$	−0.299	−0.295	−0.295
33	3-$O(CH_2)_3C_6H_5$	−0.296	−0.290	−0.290
34	3-$(CH_2)_2$-3-C_5H_4N	−0.294	−0.298	−0.298
35	3-$(CH_2)_4C_6H_4$-4-NHAc	−0.294	−0.281	−0.282
36	3-$(CH_2)_2$-2-C_5H_4N	−0.291	−0.300	−0.299
37	3-$O(CH_2)_3OC_6H_4$-2-NH_2	−0.283	−0.288	−0.288
38	3-$O(CH_2)_3OC_6H_4$-4-NHAc	−0.278	−0.270	−0.270
39	3-$(CH_2)_4$-3-C_5H_4N	−0.276	−0.284	−0.284
40	3-$O(CH_2)_4C_6H_5$	−0.276	−0.277	−0.277
41	3-$O(CH_2)_3OC_6H_4$-3-NHAc	−0.270	−0.260	−0.260
42	3-$O(CH_2)_3OC_6H_3$-3,4-Cl_2	−0.265	−0.271	−0.271
43	3-$O(CH_2)_3OC_6H_4$-3-NH_2	−0.265	−0.283	−0.283
44	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_4$-4-SO_2F	−0.265	−0.247	−0.247
45	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_5$	−0.265	−0.258	−0.258
46	3-$O(CH_2)_3OC_6H_4$-4-OCH_3	−0.262	−0.275	−0.274
47	3-$O(CH_2)_4OC_6H_4$-4-$NHCONHC_6H_4$-4-SO_2F	−0.260	−0.236	−0.237
48	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_3$-2-OCH_3-5-SO_2F	−0.260	−0.226	−0.227
49	3-$O(CH_2)_3OC_6H_4$-4-Cl	−0.257	−0.287	−0.286
50	3-$O(CH_2)_3OC_6H_4$-2-NO_2	−0.257	−0.279	−0.279
51	3-$O(CH_2)_3OC_6H_4$-3-NO_2	−0.257	−0.268	−0.268
52	3-$O(CH_2)_3OC_6H_4$-3-OCH_3	−0.256	−0.255	−0.255
53	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_3$-2-Cl-6-SO_2F	−0.255	−0.247	−0.248
54	3-$O(CH_2)_3OC_6H_4$-2-$NHCONHC_6H_5$	−0.255	−0.260	−0.259
55	3-$O(CH_2)_2OC_6H_4$-2-$NHCONHC_6H_3$-2-Cl-5-SO_2F	−0.250	−0.246	−0.246

Table 1 *(Continued)*

No.	X	1/log C		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
56	$3-O(CH_2)_3OC_6H_4-2-NHCONHCH_2C_6H_4-4-SO_2F$	−0.250	−0.232	−0.232
57	$3-O(CH_2)_3OC_6H_4-2-NHCONH-C_6H_2-2,4-(CH_3)_2-5-SO_2F$	−0.248	−0.242	−0.242
58	$3-O(CH_2)_3OC_6H_4-4-COOCH_3$	−0.247	−0.261	−0.261
59	$3-O(CH_2)_3OC_6H_3-3-NO_2-4-CH_3$	−0.245	−0.268	−0.267
60	$3-O(CH_2)_3OC_6H_4-3-CF_3$	−0.245	−0.276	−0.275
61	$3-O(CH_2)_3OC_6H_4-2-NHCONHC_6H_4-4-CH_3-3-SO_2F$	−0.245	−0.232	−0.232
62	$3-O(CH_2)_3OC_6H_4-4-NHCOC_6H_5$	−0.244	−0.242	−0.242
63	$3-O(CH_2)_3OC_6H_4-2-NHCOCH_2OC_6H_4-4-SO_2F$	−0.244	−0.239	−0.239
64	$3-O(CH_2)_3OC_6H_4-4-NHCOC_6H_4-4-OCH_3$	−0.243	−0.228	−0.229
65	$3-O(CH_2)_3OC_6H_4-2-NHCOC_6H_4-3-SO_2F$	−0.243	−0.234	−0.234
66	$3-O(CH_2)_3OC_6H_4-2-NHCOCH_2C_6H_4-4-SO_2F$	−0.243	−0.242	−0.242
67	$3-O(CH_2)_3OC_6H_4-3-COOCH_3$	−0.242	−0.256	−0.256
68	$3-O(CH_2)_3OC_6H_4-2-NHCO(CH_2)_2C_6H_4-4-SO_2F$	−0.242	−0.232	−0.232
69	$3-O(CH_2)_3OC_6H_4-4-NHCOC_6H_4-4-NO_2$	−0.239	−0.234	−0.234
70	$3-O(CH_2)_3OC_6H_4-2-NHCOC_6H_4-4-NO_2$	−0.239	−0.248	−0.248
71	$3-O(CH_2)_3OC_6H_4-4-NHCONHC_6H_5$	−0.237	−0.252	−0.252
72	$3-O(CH_2)_3OC_6H_4-4-NHCOC_6H_4-3-NO_2$	−0.237	−0.225	−0.225
73	$3-O(CH_2)_3OC_6H_4-2-NHCO(CH_2)_4C_6H_4-4-SO_2F$	−0.237	−0.220	−0.221
74	$3-O(CH_2)_3OC_6H_4-2-NHCONHC_6H_4-4-SO_2F$	−0.237	−0.248	−0.248
75	$3-O(CH_2)_3OC_6H_4-3-NHCONHC_6H_4-4-SO_2F$	−0.236	−0.231	−0.231
76	$3-O(CH_2)_3OC_6H_4-2-NHCONH(CH_2)_2C_6H_4-4-SO_2F$	−0.236	−0.224	−0.224
77	$3-O(CH_2)_4OC_6H_4-3-NHCOC_6H_4-4-SO_2F$	−0.236	−0.222	−0.222
78	$3-O(CH_2)_3OC_6H_4-2-NHCONHC_6H_3-4-Cl-3-SO_2F$	−0.235	−0.236	−0.236
79	$3-O(CH_2)_4OC_6H_4-2-NHCOC_6H3-4-CH3-3-SO_2F$	−0.235	−0.229	−0.229
80	$3-O(CH_2)_3OC_6H_4-2-NHCOC_6H_2-2,4-(CH_3)_2-5-SO_2F$	−0.234	−0.238	−0.237
81	$3-O(CH_2)_3OC_6H_4-2-NHCOC_6H_2-2,4-Cl_2-5-SO_2F$	−0.234	−0.243	−0.243
82	$3-(CH2)_4C_6H_4-2-NHCONHC_6H_4-3-SO_2F$	−0.234	−0.247	−0.246
83	$3-O(CH_2)_3OC_6H_4-3-NHCOC_6H_4-4-OCH_3$	−0.233	−0.219	−0.219
84	$3-(CH2)_4C_6H_4-2-NHCONHC_6H_4-4-SO_2F$	−0.233	−0.263	−0.261
85	$3-O(CH_2)_3OC_6H_4-4-NHCOC_6H_4-4-Cl$	−0.232	−0.238	−0.238
86	$3-O(CH_2)_3OC_6H_4-2-NHCOC_6H_3-2-CH_3-5-SO_2F$	−0.232	−0.234	−0.234
87	$3-O(CH_2)_4OC_6H_4-4-NHCONHC_6H_3-2-OCH_3-5-SO_2F$	−0.232	−0.214	−0.215
88	$3-O(CH_2)_3OC_6H_4-4-C_6H_5$	−0.230	−0.256	−0.254
89	$3-O(CH_2)_3OC_6H_4-2-NHCONHC_6H_4-3-SO_2F$	−0.230	−0.232	−0.232
90	$3-O(CH_2)_3OC_6H_4-3-NHCOC_6H_4-3-SO_2F$	−0.230	−0.210	−0.211
91	$3-O(CH_2)_2OC_6H_4-3-NHCOC_6H_4-3-SO_2F$	−0.229	−0.222	−0.222
92	$3-O(CH_2)_3OC_6H_4-4-CH_3-3-NHCOC_6H_4-4-SO_2F$	−0.229	−0.227	−0.227
93	$3-O(CH_2)_3OC_6H_4-3-NHCONHC_6H_4-3-SO_2F$	−0.222	−0.216	−0.216
94	$3-O(CH_2)_3OC_6H_4-3-NHCOCH_2C_6H_4-4-SO_2F$	−0.220	−0.222	−0.222
95	$3-O(CH_2)_3OC_6H_4-3-NHCOC_6H_4-4-SO_2F$	−0.219	−0.224	−0.224
96	$3-O(CH_2)_3OC_6H_4-2-NHCONHC_6H_3-2-Cl-5-SO_2F$	−0.217	−0.235	−0.235
97	$3-O(CH_2)_3OC_6H_4-3-NHCOCH_2OC_6H_4-4-SO_2F$	−0.217	−0.218	−0.218
98	$3-O(CH_2)_2OC_6H_4-3-NHCONHC_6H_4-4-SO_2F$	−0.216	−0.245	−0.244
99	$3-O(CH_2)_4OC_6H_4-3-NHCONHC_6H_4-4-SO_2F$	−0.215	−0.229	−0.229
100	$3-O(CH_2)_3OC_6H_4-3-NHCOC_6H_4-4-NO_2$	−0.214	−0.226	−0.226
101	$3-O(CH_2)_2OC_6H_4-3-NHCOC_6H_4-4-SO_2F$	−0.214	−0.238	−0.237
102	$3-O(CH_2)_4OC_6H_4-2-NHCONHC_6H_3-2-Cl-5-SO_2F$	−0.207	−0.231	−0.231
103	$3-O(CH_2)_3OC_6H_4-3-NHCONHC_6H_4-4-NO_2$	−0.204	−0.233	−0.232
104	$3-O(CH_2)_3OC_6H_4-4-CH_3-3-NHCONHC_6H_4-4-SO_2F$	−0.204	−0.224	−0.223
105	$3-O(CH_2)_3OC_6H_4-3-NHCONH(CH_2)_2C_6H4-4-SO_2F$	−0.193.	−0.203	−0.203

[a] CV and FIT values are calculated using Eq. (8).

Table 2
69 benzene derivatives and their observed and calculated (cross-validated, CV, and fitted, FIT) fathead minnow toxicities, expressed as $-\log(LC_{50})$

No.	Compound	$-\log(LC_{50})$		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
1	Benzene	3.40	3.29	3.32
2	Bromobenzene	3.89	4.04	4.01
3	Chlorobenzene	3.77	3.75	3.75
4	Phenol	3.51	3.31	3.35
5	Toluene	3.32	3.51	3.49
6	1.2-Dichlorobenzene	4.40	4.33	4.33
7	1.3-Dichlorobenzene	4.30	4.10	4.12
8	1.4-Dichlorobenzene	4.62	4.80	4.77
9	2-Chlorophenol	4.02	4.01	4.01
10	3-Chlorotoluene	3.84	3.72	3.73
11	4-Chlorotoluene	4.33	4.11	4.13
12	1.3-Dihydroxybenzene	3.04	3.31	3.28
13	3-Hydroxyanisole	3.21	3.13	3.14
14	2-Methylphenol	3.77	3.62	3.62
15	3-Methylphenol	3.29	3.52	3.51
16	4-Methylphenol	3.58	3.64	3.64
17	4-Nitrophenol	3.36	3.68	3.66
18	1.4-Dimethoxybenzene	3.07	3.01	3.01
19	1.2-Dimethylbenzene	3.48	3.84	3.81
20	1.4-Dimethylbenzene	4.21	3.94	3.97
21	2-Nitrotoluene	3.57	3.70	3.69
22	3-Nitrotoluene	3.63	3.67	3.66
23	4-nitrotoluene	3.76	3.71	3.71
24	1.2-Dinitrobenzene	5.45	4.95	5.09
25	1.3-Dinitrobenzene	4.38	4.12	4.15
26	1.4-Dinitrobenzene	5.22	4.83	4.91
27	2-Methyl-3-nitroaniline	3.48	3.74	3.73
28	2-Methyl-4-nitroaniline	3.24	3.50	3.47
29	2-Methyl-5-nitroaniline	3.35	3.80	3.77
30	2-Methyl-6-nitroaniline	3.80	3.76	3.76
31	3-Methyl-6-nitroaniline	3.80	3.61	3.62
32	4-Methyl-2-nitroaniline	3.79	3.78	3.78
33	4-Hydroxy-3-nitroaniline	3.65	3.51	3.52
34	4-Methyl-3-nitroaniline	3.77	3.78	3.78
35	1,2,3-Trichlorobenzene	4.89	4.84	4.84
36	1,2,4-Trichlorobenzene	5.00	5.02	5.02
37	1,3,5-Trichlorobenzene	4.74	4.36	4.45
38	2,4-Dichlorophenol	4.30	4.53	4.52
39	3,4-Dichlorotoluene	4.74	4.46	4.48
40	2,4-Dichlorotoluene	4.54	4.57	4.56
41	4-Chloro-3-methylphenol	4.27	4.27	4.27
42	2,4-Dimethylphenol	3.86	3.74	3.76
43	2,6-Dimethylphenol	3.75	3.75	3.75
44	3,4-Dimethylphenol	3.90	3.90	3.90
45	2,4-Dinitrophenol	4.04	4.03	4.04
46	1,2,4-Trimethylbenzene	4.21	4.07	4.09
47	2,3-Dinitrotoluene	5.01	5.29	5.21
48	2,4-Dinitrotoluene	3.75	4.29	4.27
49	2,5-Dinitrotoluene	5.15	4.89	4.93
50	2,6-Dinitrotoluene	3.99	4.43	4.41
51	3,4-Dinitrotoluene	5.08	5.29	5.23
52	3,5-Dinitrotoluene	3.91	4.25	4.23
53	1,3,5-Trinitrobenzene	5.29	5.29	5.29
54	2-Methyl-3,5-dinitroaniline	4.12	4.23	4.22

Table 2 *(Continued)*

No.	Compound	$-\log(LC_{50})$		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
55	2-Methyl-3,6-dinitroaniline	5.34	4.59	4.64
56	3-Methyl-2,4-dinitroaniline	4.26	3.97	4.00
57	5-Methyl-2,4-dinitroaniline	4.92	3.88	3.97
58	4-Methyl-2,6-dinitroaniline	4.21	4.76	4.72
59	5-Methyl-2,6-dinitroaniline	4.18	4.64	4.61
60	4-Methyl-3,5-dinitroaniline	4.46	4.33	4.34
61	2,4,6-Tribromophenol	4.70	4.98	4.82
62	1,2,3,4-Tetrachlorobenzene	5.43	5.55	5.53
63	1,2,4,5-Tetrachlorobenzene	5.85	5.76	5.77
64	2,4,6-Trichlorophenol	4.33	4.68	4.64
65	2-Methyl-4,6-dinitrophenol	5.00	4.45	4.48
66	2,3,6-Trinitrotoluene	6.37	6.39	6.38
67	2,4,6-Trinitrotoluene	4.88	5.32	5.26
68	2,3,4,5-Tetrachlorophenol	5.72	5.64	5.65
69	2,3,4,5,6-Pentachlorophenol	6.06	6.01	6.03

[a] CV and FIT values are calculated using Eq. (10).

which was achieved by the orthogonalization of descriptors, because in the orthogonal basis the computation of R is much faster and simpler (Lučić et al., 1995a,b,c; Lučić, 1997). Namely, in the case one has the MR model based on the set of I orthogonalized descriptors di ($i = 1, \ldots, I$), the correlation coefficient between the experimental values of modeled activity A and the values estimated by the model A^{est} can be calculated in a very simple way (Eq. (1)):

$$R = \left[\sum_{i=1}^{I} R_i^2 \right]^{1/2} \tag{1}$$

where Ri is the correlation coefficient between each orthogonalized descriptor di and the modeled activity A. For example, using this procedure it takes 28 CPU min on Hewlett-Packard 9000/E55 computer, which is configured as a server, to select the best MR model with five out of 104 descriptors among $\sim 10^8$ possible models.

3. Results and discussion

3.1. QSAR of benzamidines

The best one-descriptor structure–complement–inhibitory activity model of benzamidines obtained is:

$$1/\log C = -0.9332(\pm 0.0229) + 0.4395(\pm 0.0152) H^V$$
$$n = 105 \; R = 0.943 \; R_{cv} = 0.941 \; S = 0.0195 \; S_{cv} = 0.0199 \; F = 832 \tag{2}$$

where H^V is the graph-vertex complexity (Basak, 1987), n is the number of benzamidine derivatives considered, R is the correlation coefficient, R_{cv} is the leave-one-out (cross-validated) correlation coefficient, F is F-value, S is the standard error and S_{cv} is the cross-validated (leave-one-out) standard error of estimate (root-mean-square error), both with N-2 in the denominator. This model is only slightly better than the earlier obtained one-descriptor model, but with a different descriptor (Basak et al., 1999a):

$$1/\log C = -0.6428(\pm 0.0129) + 0.0490(\pm 0.0017)^{3D}W$$
$$n = 105 \; R = 0.943 \; R_{cv} = 0.940 \; S = 0.0196 \; S_{cv} = 0.0200 \; F = 824 \tag{3}$$

where ^{3D}W is the 3-D Wiener number for the hydrogen-suppressed structures computed using their geometric distance matrices (Bogdanov et al., 1989). Close to this model is a model with 3-D Wiener number computed for structures containing all atoms including hydrogens (Bosnjak et al., 1991) ($n = 105$, $R = 0.941$, $R_{cv} = 0.939$, $S = 0.0199$ $S_{cv} = 0.0203$).

The best two-descriptor model of the benzamidine structure-complement-inhibitory activity is:

$$1/\log C = -0.6878(\pm 0.0175) + 0.1327(\pm 0.0367) W + 0.1864(\pm 0.0380)^{3D}W$$
$$n = 105 \; R = 0.950 \; R_{cv} = 0.947 \; S = 0.0184 \; S_{cv} = 0.0189 \; F = 467 \tag{4}$$

where W is the 2-D Wiener number (Wiener, 1947). The best three-descriptor model is given by:

Table 3

Descriptions of all considered descriptors and symbols of only those descriptors involved in the models

	Information index for the magnitude of distances between all possible pairs of vertices of a graph
	Mean information index for the magnitude of distance
W	Wiener index, the half-sum of the off-diagonal elements of the molecular distance matrix
	Degree complexity
H^V	Graph vertex complexity
	Graph distance complexity
	Information content of the distance matrix partitioned by frequency of occurrences of distance l
	Information content of the hydrogen-suppressed graph at its maximum neighborhood of vertices
	Order of neighborhood when ICr reaches its maximum value for the hydrogen-filled graph
	A Zagreb group parameter, the sum of square of degree over all vertices
	A Zagreb group parameter, the sum of cross-product of degrees over all neighboring (connected) vertices
Ic_r	Mean information content of a graph based on the rth ($r = 0$–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth ($r = 0$–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for rth ($r =$ 0–6) order neighborhood of vertices in a hydrogen-filled graph
	Path connectivity index of order $h = 0$–6
	Cluster connectivity index of order $h = 3$–6
	Chain connectivity index of order $h = 6$
	Path-cluster connectivity index of order $h = 4$–6
	Bond path connectivity index of order $h = 0$–6
${}^h\chi^b_c$	Bond cluster connectivity index of order $h = 3$–6
${}^h\chi^b_{ch}$	Bond chain connectivity index of order $h = 6$
	Bond path-cluster connectivity index of order $h =$ 4–6
${}^h\chi^v$	Valence path connectivity index of order $h = 0$–6
${}^h\chi^v_c$	Valence cluster connectivity index of order $h = 3$–6
${}^h\chi^v_{ch}$	Valence chain connectivity index of order $h = 6$
${}^h\chi^v_{Pc}$	Valence path-cluster connectivity index of order $h = 4$–6
P_l	Number of paths of length $l = 0$–10
	Balaban's J index based on distance
	Balaban's J index based on relative electronegativities
	Balaban's J index based on relative covalent radii
	Balaban's J index based on bond types
	Energy of the highest occupied molecular orbital
	Energy of the second highest occupied molecular orbital
E_{lumo}	Energy of the lowest unoccupied molecular orbital
	Energy of the second lowest unoccupied molecular orbital
ΔH_f	Heat of formation
μ	Dipole moment
	Van der WaalSs volume
${}^{3D}W_H$	3-D Wiener index for the hydrogen-filled geometric distance matrix
${}^{3D}W$	3-D Wiener index for the hydrogen-suppressed geometric distance matrix

$$1/\log C = -0.6400(\pm 0.0239) + 0.1273(\pm 0.0355)W + 0.0103(\pm 0.0037)P_9 + 0.1698(\pm 0.0372)^{3D}W$$

$$n = 105 \quad R = 0.954 \quad R_{cv} = 0.949 \quad S = 0.0177 \quad S_{cv} = 0.0185 \quad F = 335 \qquad (5)$$

where P_9 is the path of length nine. P_9 could be omitted from Eq. (5) because the related value of error of regression coefficient is relatively large comparing to the value of regression coefficient. Then Eq. (5) simply converts into Eq. (4). The best four-descriptor model is:

$$1/\log C = -0.6999(\pm 0.0194) + 0.1327(\pm 0.0354)W + 5.0332(\pm 1.2285)^6\chi^b_{ch} - 5.1120\ (\pm 1.2486)^6\chi^v_{ch} + 0.1885(\pm 0.0359)^{3D}W$$

$$n = 105 \quad R = 0.957 \quad R_{cv} = 0.953 \quad S = 0.0170 \quad S_{cv} = 0.0177 \quad F = 272 \qquad (6)$$

where ${}^6\chi^b_{ch}$ and ${}^6\chi^v_{ch}$ denote the bond-chain and valence-chain connectivity indices of order six, respectively.

Hansch and Yoshimoto (Hansch and Yoshimoto, 1974) published, 25 years ago, the following four-descriptor model for benzamidine derivatives inhibiting complement (the model is given in their notation):

$$\log(1/C) = 0.15(\pm 0.03)(\mathrm{MR} - 1.2) + 1.07(\pm 0.13)(\text{D-1}) + 0.52(\pm 0.28)(\text{D-2}) + 0.43(\pm 0.14)(\text{D-3}) + 2.425(\pm 0.12)$$

$$n = 108 \quad R = 0.935 \quad S = 0.258 \qquad (7)$$

where MR is the molar refractivity of substituents at positions 1 and 2, taken from the compilation by Hansch et al. (Hansch et al., 1973) or computed, while D-1, D-2, and D-3 are indicator variables for the presence or absence of three kinds of the substructural units in a given benzadimine. To compare fitted statistical parameters of our four-descriptor model (Eq. (6)) with those of model given by Eq. (7), we retransformed our results into a log (1/C) scale used by Hansch and Yoshimoto. Thus, we obtained statistical parameters ($R = 0.941$ and $S = 0.237$) that are comparable with their result. However, Hansch and Yoshimoto considered 108 benzamidine derivatives and we only considered 105. This discrepancy is caused by problematic data for three compounds which in our case are discarded from the set of benzamidine derivatives (Basak et al., 1999a). But, the nature of descriptors used in these two types of models is different. Descriptors used by us are calculated solely from the structures of studied molecules while the Hansch–Yoshimoto parameters (molar refractivities of substituents) are experimentally-based.

Finally, the five-descriptor model is:

$$
\begin{aligned}
1/\log C = {} & 1.5264(\pm 0.3534) + 0.6323(\pm 0.0936)(\mathrm{IC})_2 \\
& - 1.6788(\pm 0.2720)(\mathrm{IC})_6 \\
& - 1.4540(\pm 0.2043)(\mathrm{SIC})_1 \\
& - 0.4239(\pm 0.0680)(\mathrm{CIC})_6 \\
& + 0.1286(\pm 0.0149)^{3\mathrm{D}}W
\end{aligned}
$$

$$
n = 105 \ R = 0.963 \ R_{\mathrm{cv}} = 0.957 \ S = 0.0158 \ S_{\mathrm{cv}} = 0.0170 \ F = 253 \qquad (8)
$$

where $(\mathrm{IC})_2$ and $(\mathrm{IC})_6$ denote the mean information content of structure based on the second- and sixth-order neighborhood of atoms, including hydrogens, in the structure, respectively, $(\mathrm{SIC})_1$ and $(\mathrm{CIC})_6$ are, respectively, the structural information content for the first order neighborhood and complementary information content for the sixth order neighborhood of atoms, including hydrogens, in the structure. $(\mathrm{IC})_r$, $(\mathrm{SIC})_r$ and $(\mathrm{CIC})_r$ are molecular complexity indices introduced some times ago by one of us (Basak, 1987) for use in predictive pharmacology and toxicology.

It is interesting to note that the 3-D Wiener number is present in all models given here, except in the very best model with a single descriptor, although is present in the next best single-descriptor model. This is not surprising because this descriptor has shown to be very useful in the structure–property–activity modeling (Bogdanov et al., 1989; Bosnjak et al., 1991; Mihalić and Trinajstić, 1991; Nikolić et al., 1991; Trinajstić, 1992).

The models containing more decriptors did not outperform the above five-descriptor model. Thus, the model with five-descriptors (Eq. (8)), selected from the initial set of descriptors, is the best QSAR model, according to the calculated cross-validated statistical parameters, for predicting the benzamidine structure–complement–inhibitory activity. This model is better than one-descriptor model previously obtained using hierarchical approach (Basak et al., 1999a). However,

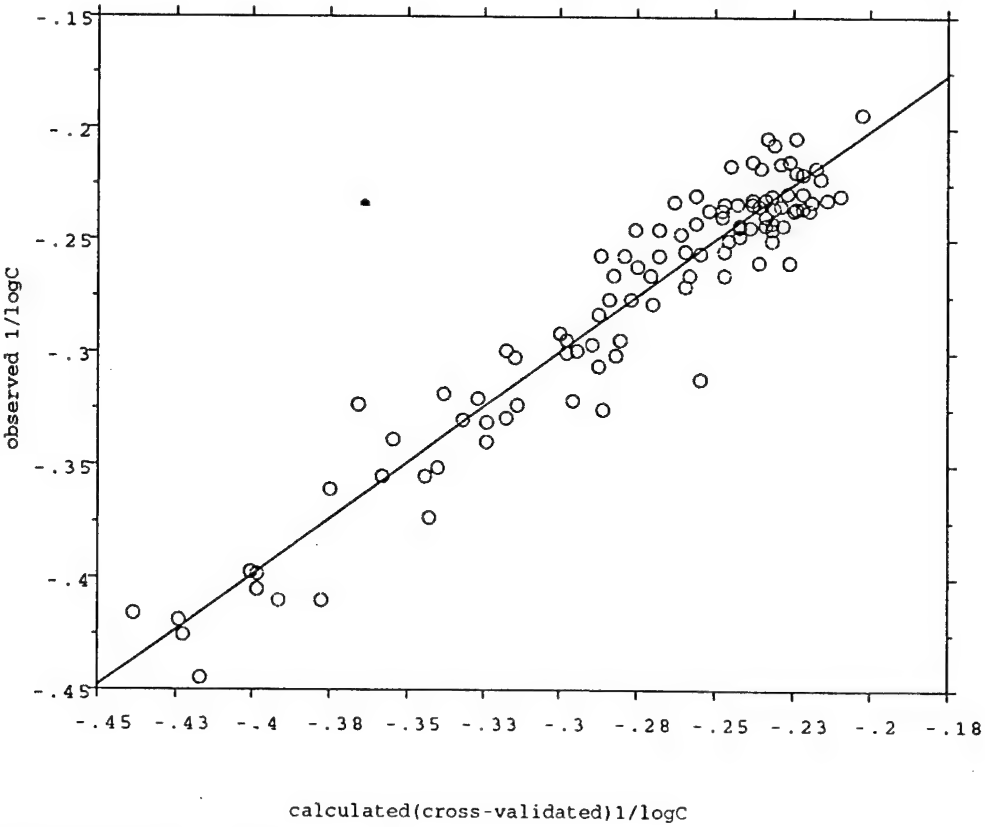


Fig. 2. A plot of observed versus calculated (cross-validated) 1/log C complement–inhibitory activity of benzamidines.

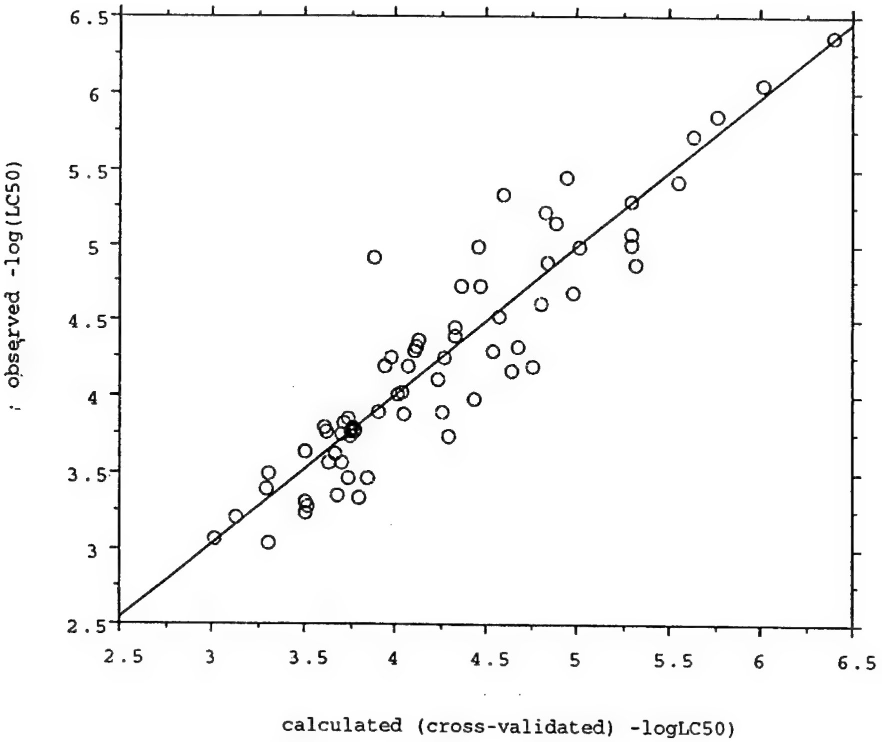


Fig. 3. A plot of observed versus calculated (cross-validated) $-\log(\mathrm{LC}_{50})$ benzene derivatives acute toxicities.

according to F-values one-descriptor models selected in this paper and our previous work (Basak et al., 1999a) appear to be better models than the model with five-descriptors. But, the F-value is calculated only from the fitted correlation coefficient R and taking into account the number of parameters optimized in the model. Because it is accepted (Ortiz et al., 1997) that the cross-validated statistical parameters give better evidence into the model quality than fitted statistical parameters, our final conclusions are based on cross-validated statistical parameters, although the prediction for compounds from an external data set would be the best way of model quality testing. A plot between the experimental and predicted values, calculated in the cross-validation procedure using Eq. (8), of 1/log C is given in Fig. 2. Computed (fitted and leave-one-out cross-validated) 1/log C values are given in Table 1.

3.2. QSAR of benzene derivatives

The best linear five-descriptor structure–toxicity model of benzene derivatives selected by CROMRsel program is:

$$
\begin{aligned}
&-\log(\mathrm{LC}_{50})\\
&= 5.2032(\pm 0.546) + 0.8488(\pm 0.106)P_9\\
&\quad + 1.7979(\pm 0.183)^4\chi^{\mathrm{v}}_{\mathrm{Pc}} - 0.4439(\pm 0.0523)E_{\mathrm{lumo}}\\
&\quad - 0.1379(\pm 0.0195)\mu - 0.2961(\pm 0.0927)^{\mathrm{3D}}W\mathrm{H}\\
&n = 69 \;\; R = 0.927 \;\; R_{\mathrm{cv}} = 0.914 \;\; S = 0.287 \;\; S_{\mathrm{cv}} = 0.312\\
&F = 77 \qquad (9)
\end{aligned}
$$

where P_9 is the path of length nine, $^4\chi^{\mathrm{v}}_{\mathrm{Pc}}$ valence path-cluster connectivity index of order four, E_{lumo} is the energy of the lowest unoccupied molecular orbital, μ is dipole moment, and $^{\mathrm{3D}}W_{\mathrm{H}}$ is the 3-D Wiener number for the hydrogen-filled structures computed using their geometric distance matrices (Bogdanov et al., 1989). This model has two descriptors fewer than the best model obtained by hierarchical approach (see Gute and Basak, 1997) and possesses almost the same statistical parameters.

The best linear seven-descriptor model is:

$$
\begin{aligned}
&-\log(\mathrm{LC}_{50})\\
&= 4.4100(\pm 0.809) + 0.8637(\pm 0.0988)P_9\\
&\quad + 2.5278(\pm 0.833)^2\chi^{\mathrm{v}} - 3.1248(\pm 0.655)^4\chi^{\mathrm{v}}\\
&\quad + 1.5628(\pm 0.372)^6\chi^{\mathrm{v}}_{\mathrm{Pc}} - 0.44157(\pm 0.051)E\mathrm{lumo}\\
&\quad - 0.1364(\pm 0.018)\mu - 0.34054(\pm 0.087)^{\mathrm{3D}}W\mathrm{H}\\
&n = 69 \;\; R = 0.940 \;\; R_{\mathrm{cv}} = 0.925 \;\; S = 0.262 \;\; S_{\mathrm{cv}} = 0.291 \;\; F\\
&= 66 \qquad (10)
\end{aligned}
$$

where ${}^{2}\chi^{v}$ and ${}^{4}\chi^{v}$ denote valence path connectivity indices of order two and four, respectively, and ${}^{6}\chi^{v}_{Pc}$ is the valence path-cluster connectivity index of order six. Other descriptors are the same as those from five-descriptor model (Eq. (9)). This model ($R^2 = 0.884$, $F = 66$, $S = 0.26$) is better than the seven-descriptor model obtained by hierarchical procedure (see Gute and Basak, 1997) ($R^2 = 0.863$, $F = 50$, $S = 0.30$), and one can see that these two models contain three identical descriptors: P_9, ${}^{3D}W_H$, and μ. Fitted and cross-validated predicted values for all benzene derivatives obtained using Eq. (10) are given in Table 2. A plot between the experimental and predicted values, calculated in the cross-validation procedure using Eq. (10), of $-\log(LC_{50})$ is given in Fig. 3.

We also found several seven-descriptor linear multiregression models with better statistical prameter than the best seven-descriptor model of Gute and Basak (see Gute and Basak, 1997). One of them is very similar to the model given as Eq. (10) and involving the following set of descriptors H^v, P_9, ${}^{3}\chi^{b}_{c}$, ${}^{5}\chi^{v}_{c}$, ΔH_f, μ, ${}^{3D}W_H$ (see Table 3 for description of descriptors), and possessing the following statistical parameters $R = 0.9398$, $R_{cv} = 0.9245$, $S = 0.262$, $S_{cv} = 0.292$, $F = 66$).

In addition, we perform modeling in order to compare our seven-descriptor model with the additivity model (using eight terms, i.e. eight optimized parameters) derived by Hall et al. (Hall et al., 1984). To do this we omitted from the data set compounds 53, 57 and 65, which were identified in by Hall et al. as outliers. For 66 compounds statistical parameters of seven-descriptor model (Eq. (10)) are: $R = 0.955$, $R_{cv} = 0.943$, $S = 0.225$, $S_{cv} = 0.255$ $F = 87$). This parameters are better than those for additivity models obtained by Hall et al. ($R = 0.951$, $S = 0.249$, $F = 67$).

4. Concluding remark

Presented results show that the optimum way to carry out QSAR modeling is by selecting the best descriptors in (linear, as was the case here, or nolinear (Lučić and Trinajstić, 1999) multiregression models.

Acknowledgements

This is contribution number 256 from the Center for Water and the Environment of the Natural Resources Research Institute. This work was supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force and 00980606 from the Ministry of Science and Technology of the Republic of Croatia. We thank reviewers for helpful comments.

References

Basak, S.C., 1987. Use of molecular complexity indices in predictive pharmacology and toxicology. Med. Sci. Res. 15, 605–609.

Basak, S.C., Gute, B.D., Ghatak, S., 1999a. Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. J. Chem. Inf. Comput. Sci. 39, 255–260.

Basak, S.C., Gute, B.D., Grunwald, G.D., 1997. Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. J. Chem. Inf. Comput. Sci. 37, 651–655.

Basak, S.C., Gute, B.D., Grunwald, G.D., 1999b. In: Devillers, J., Balaban, A.T. (Eds.), Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Reading, pp. 245–261.

Basak, S.C., Gute, B.D., Opitz, D.W., Balasubramanian, K., 1999c. Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach. Reported at the American Association of Artificial Intelligence (AAAI) Conference — Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, Stanford University, March 22–24.

Bogdanov, B., Nikolić, S., Trinajstić, N., 1989. On the three-dimensional Wiener number. J. Math. Chem. 3, 299–309.

Bosnjak, N., Mihalić, Z., Trinajstić, N., 1991. Application of topographic indices to chromatographic data: calculation of the retention indices of alkanes. J. Chromatogr. 540, 430–440.

Free, S.M., Wilson, J.W., 1964. A mathematical contribution to structure–activity studies. J. Med. Chem. 1, 395–399.

Gute, B.D., Basak, S.C., 1997. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. SAR QSAR Environ. Res. 7, 117–131.

Hall, L.H., Kier, L.B., Phipps, G., 1984. Structure–activity relationship studies on the toxicities of benzene derivatives: I. an additivity model. Environ. Toxicol. Chem. 3, 355–365.

Hansch, C., Leo, A., Unger, S.H., Kim, K.H., Nikaitani, D., Lien, E.J., 1973. Aromatic substituent constants for structure–activity correlations. J. Med. Chem. 16, 1207–1216.

Hansch, C., Yoshimoto, M., 1974. Structure–activity relationships in immunochemistry. 2. Inhibition of complement by benzamidines. J. Med. Chem. 17, 1160–1167.

Katritzky, A.R., Chen, K., Wang, Y., Karelson, M., Lučić, B., Trinajstić, N., Suzuki, T., Schüürmann, G., 1999. Prediction of liquid viscosity for organic compounds by a quantitative structure–property relationship. J. Phys. Org. Chem. (in press).

Kubinyi, H., Kehrhahn, O.H., 1976. Quantitative structure–activity relationships: a comparison of different Free-Wilson models. J. Med. Chem. 19, 1040–1045.

Kuby, J., 1992. Immunology. Freeman, New York.

Lučić, B., 1997. Ph. Dissertation, University of Zagreb, Zagreb. CROMRsel.f (CROatian MultiRegression selection of descriptors) is a computer program for the selection of descriptors for the best MR models.

Lučić, B., Amić, D., Trinajstić, N., 1999a. Nonlinear multivariate regression outperforms several concisely designed neural networks in QSAR modeling. J. Chem. Inf. Comput. Sci. (in press).

Lučić, B., Nikolić, S., Trinajstić, N., Juretić, D., 1995a. The structure–property models can be improved using the orthogonalized descriptors. J. Chem. Inf. Comput. Sci. 35, 532–538.

Lučić, B., Nikolić, S., Trinajstić, N., Juretić, D., Jurić, A., 1995b. A novel QSPR approach to physicochemical properties of the α-amino acids. Croat. Chem. Acta 68, 435–450.

Lučić, B., Nikolić, S., Trinajstić, N., Jurić, A., Mihalić, Z., 1995c. A structure–property study of the solubility of aliphatic alcohols in water. Croat. Chem. Acta 68, 417–434.

Lučić, B., Trinajstić, N., 1999. Multivariate regression outperforms several robust architectures of neural networks in QSAR modeling. J. Chem. Inf. Comput. Sci. 39, 121–132.

Lučić, B., Trinajstić, N., Sild, S., Karelson, M., Katritzky, A.R., 1999b. A new efficient approach for variable selection based on multiregression: rediction of gas chromatographic retention times and response factors. J. Chem. Inf. Comput. Sci. 39, 610–621.

Mihalić, Z., Trinajstić, N., 1991. The algebraic modelling of chemical structures: on the development of three-dimensional molecular descriptors. J. Mol. Struct. (Theochem.) 232, 65–78.

Nikolić, S., Trinajstić, N., Mihalić, Z., Carter, S., 1991. On the geometric distance matrix and the corresponding structural invariants of molecular systems. Chem. Phys. Lett. 179, 21–28.

Ortiz, A.R., Pastor, M., Palomer, A., Cruciani, G., Gago, F., Wade, R.C., 1997. Reliability of comparative molecular field analysis models: effect of data scaling and variable selection using a set of human synovial fluid phospholipase A_2 inhibitors. J. Med. Chem. 40, 1136–1148.

Trinajstić, N., 1992. Chemical Graph Theory, second revised. CRC, Boca Raton, FL, pp. 262–269.

Wiener, H., 1947. Structural determination of paraffin boiling points. J. Am. Chem. Soc. 69, 17–20.

Nikolić, S., Trinajstić, N., Mihalić, Z., Carter, S., 1991. On the geometric distance matrix and the corresponding structural invariants of molecular systems. Chem. Phys. Lett. 179, 21–28.

Ortiz, A.R., Pastor, M., Palomer, A., Cruciani, G., Gago, F., Wade, R.C., 1997. Reliability of comparative molecular field analysis models: effect of data scaling and variable selection using a set of human synovial fluid phospholipase A_2 inhibitors. J. Med. Chem. 40, 1136–1148.

Trinajstić, N., 1992. Chemical Graph Theory, second revised. CRC, Boca Raton, FL, pp. 262–269.

Wiener, H., 1947. Structural determination of paraffin boiling points. J. Am. Chem. Soc. 69, 17–20.

APPENDIX 1.7 Construction of high-quality structure-property-activity regressions: The boiling points of sulfides

Construction of High-Quality Structure−Property−Activity Regressions: The Boiling Points of Sulfides

Milan Randić*[†,‡] and Subhash C. Basak[§,⊥]

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311 and
Natural Resources Research Institute, The University of Minnesota, 5013 Miller Trunk Highway,
Duluth, Minnesota 55811

Received September 9, 1999

Instead of using the standard molecular descriptors (topological indices) for regression analysis, which are numerically fully determined once a molecule is selected, we outline the use of variable molecular descriptors that are modified during the search for the best regression. The approach is illustrated using boiling points of sulfides. We have transformed the connectivity index $^1\chi$ into a function of two variables (x, y) which differentiate carbon and sulfur atoms. The optimal values of the variables (x, y) were determined by minimizing the standard error of the regression. With the values $x = +0.25$ and $y = -0.95$ for carbon and sulfur, respectively, we have obtained a regression based on a single descriptor and a standard error of 1.8 °C. With elimination of two outliers (having a deviation of about 4 °C) the standard error is reduced to a remarkable 1.3 °C.

INTRODUCTION

The past decade has witnessed two important developments of multivariate regression analysis, MRA, relevant for quantitative structure−property-activity relationship, QSAR: (1) expansion of mathematical structural descriptors for characterization of molecular structure;[1−5] (2) construction of orthogonal molecular descriptors[6−12] which result in stable regression equations. The first, which is of interest when better regressions are sought, is rather conspicuous, while the second, which is important for interpretation of the results of such studies, remains not yet sufficiently widely appreciated.

In this paper we will address the problem of construction of high-quality regressions (HQR). With hundreds of descriptors available[13−15] the questions to consider are as follows: (1) How should an optimal set of descriptors be chosen from a large number of available descriptors? (2) How should one chose between regressions of seemingly similar quality? (3) How unique are regression results? (4) Are there important structural elements missed by the descriptors used? (5) How complete is the space spanned by molecular descriptors for the structure−property-activity studies? (6) Do we need additional molecular descriptors?

HIGH-QUALITY REGRESSIONS

The standard error in most correlations still does not approach the experimental error of measurements. How realistic is it to hope to arrive at this goal? As we will show, HQR, in which the standard error has been dramatically reduced in comparison with traditional approaches using the same number of descriptors, can be derived with a new kind of molecular descriptors which involve variability that allows one to optimize the descriptors and minimize the standard error of regression.

Table 1. Standard error for the Boiling Points of Smaller Sulfides (n = 21 Compounds) for Selection of Descriptors

descriptors	standard error	descriptors	standard error
χ, J	2.001	χ	2.701
χ, n	2.550	n, J	2.748
χ, P	2.560	n, p_2/w_2	2.981
χ, W	2.667	J, W	4.808
χ, p_2/w_2	2.692	W, P	5.109

In Table 1 we illustrate the standard errors for correlations of the boiling points of smaller sulfides (shown in Figure 1) using a selection of molecular descriptors. When the connectivity index[16] is used alone, we find the standard error of the regression is 2.70 °C, as shown in the middle of Table 1. When the connectivity index is combined with Balaban's J index,[17] the standard error is further reduced to 2.00 °C. Other descriptors, viz., n, the number of non-hydrogen atoms, P_3, the number of paths of length 3, W, the Wiener index,[18] and the p_k/w_k, path/walk quotients,[19] give only a minor improvement for the standard error over that based on $^1\chi$ used alone. In contrast other combinations of molecular descriptors (listed in the right part of Table 1) do not give satisfactory results. The standard error in such combinations is worse than the standard error when the connectivity index is used as a single descriptor, which well-illustrates the importance of the proper selection of molecular descriptors.

The compounds considered here were among 45 saturated acyclic compounds possessing divalent sulfur atoms for which Balaban et al.[20] found reliable literature data. We took all compounds having six or fewer carbon atoms, a total of 21, and have recalculated the regressions for only these smaller sulfides. The study of Balaban and co-workers considered a broader class of compounds: 185 saturated

[†] Drake University and The University of Minnesota.
[‡] FAX (home): 515 292 8629.
[§] The University of Minnesota.
[⊥] E-mail: sbasak@nrri.umn.edu.

10.1021/ci990115q CCC: $19.00 © 2000 American Chemical Society
Published on Web 03/31/2000

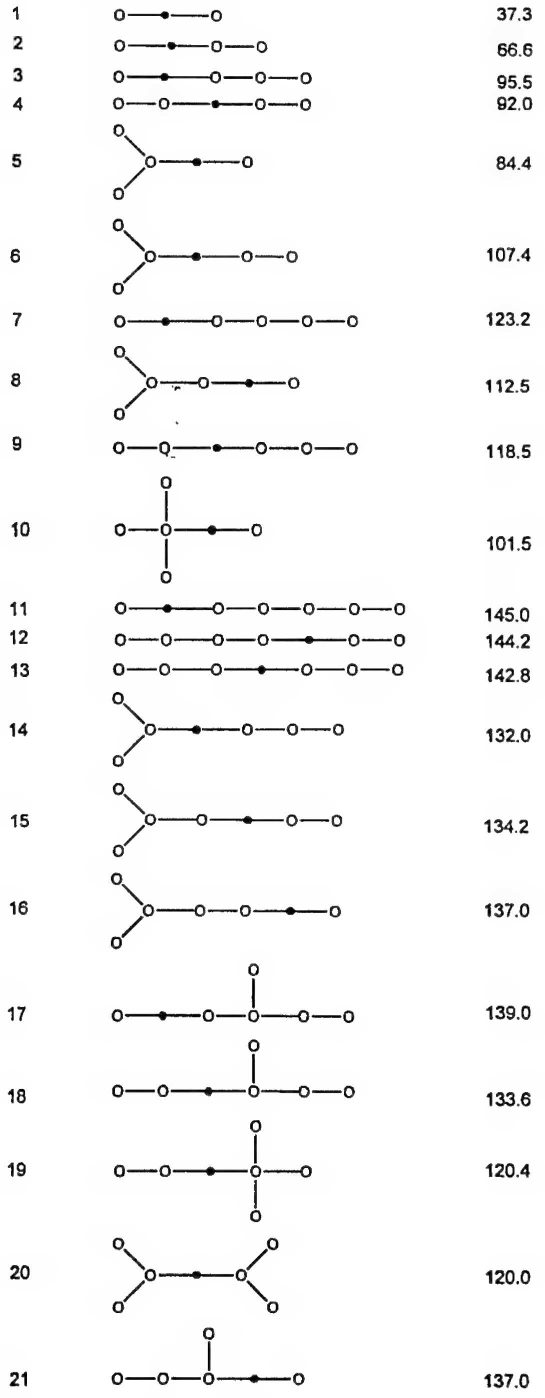


Figure 1. Molecular graphs of smaller sulfides and their boiling points. The sulfur atoms are shown as a filled circle.

acyclic compounds possessing divalent oxygen or sulfur atoms, and devoid of hydrogen bonding, having 11 or less non-hydrogen atoms. Their purpose was as follows: (i) to explore the role of heteroatoms within acyclic skeletons in determining a measured molecular property (boiling points); (ii) to show that topological descriptors can satisfactorily account for the observed relative magnitudes of the property; and (iii) to derive structure−property regressions that may be useful for predicting boiling points of unknown compounds.

Our objectives are the same, but our philosophy in this particular study is somewhat different: Rather than considering a large set of mixed compounds (alkanes, ethers, diethers, acetals, and peroxides as well as their sulfur analogues: sulfides (thioethers), bis-sulfides, thioacetals, and disulfides), which allows one to use several molecular descriptors and still maintain high statistical significance for the correlation, we decided to use only structurally closely related compounds. In particular, we excluded bis-sulfides and disulfides because of the presence of S−S linkage that is absent in sulfides. This has reduced the pool of the compounds considerably, which limits the number of descriptors that one should use in analyzing the data. By homogenizing the sample of the compounds to be examined, as we will see, we can achieve a very high quality regression result using a *single* descriptor.

As we see from Table 1, apparently it is difficult to reduce the standard error for the boiling points of sulfides below 2.5 °C. Among the combinations listed in Table 1, only Balaban's *J* reduced the standard error below 2.5 °C. This may not be surprising because all descriptors of Table 1 except *J* do not differentiate sulfur and carbon atoms. Hence, 2.5 °C may well be the limit that such models can attain. The experimental boiling points for butylmethyl sulfide (**7**) and ethylpropyl sulfide (**9**), 123.2 and 118.5 °C, respectively, differ by almost 5 °C. If we overlook the difference between sulfur and carbon, both these structures have the same molecular graph. The same is true for ethylisopropyl sulfide (**6**) and isobutylmethyl sulfide (**8**), with the boiling points 107.4 and 112.5 °C, respectively. Hence, the simple connectivity index and other topological indices that do not discriminate heteroatoms can at best approach the standard error of about 2.5 °C.

Observe that the descriptors listed in Table 1 are of quite distinct structural origin and thus do not duplicate one another. However, many of such indices, even when combined (the right part of Table 1), apparently lack flexibility to represent the data with desirable accuracy. Using descriptors that differentiate heteroatoms, we reach a standard error of about 2 °C. The question to consider is as follows: Can the standard error of 2 °C obtained using $^1\chi$ and *J* be further dramatically reduced? Have we reached the limit for correlating the boiling points of sulfides? Is it that the residual of the molecular property considered cannot be described by any of the available structural descriptors?

FLEXIBLE MOLECULAR DESCRIPTORS

In order to develop a high-quality regression, we not only need new descriptors but we need a *new kind* of molecular descriptors that have the flexibility to adjust to the variability that different molecules may show. One such descriptor has been introduced in the multiple regression analysis 10 years ago,[21,22] but apparently has been mostly overlooked. That novelty can be ignored or overlooked has already been well-illustrated by the Wiener index *W*, which waited two decades to be resurrected. In order to not repeat that history, we undertook a concerted effort to illustrate properties of variable descriptors, and the variable connectivity index, in particular.[23−26] The variable connectivity index represents an important and distinct generalization of the connectivity index $^1\chi$ since it offers a flexibility that traditional topological indices, all several hundred of them, have been lacking.

We propose here a special symbol, ${}^1\chi^f$, for the flexible connectivity, index which is to be outlined shortly. The original connectivity index ${}^1\chi$ (named so by Kier et al.[27]), proposed by Randic,[16] used a fixed number as entries in the weighting algorithm $1/(pq)^{1/2}$ for the contribution of a bond having p and q neighbors. The higher order connectivity indices, ${}^m\chi$,[28] were defined analogously using paths of length m, for m = 2, 3, ... The bonding connectivity indices, ${}^1\chi^b$, were considered by Basak and Magnuson[29] on the basis of weights equal to the number of bonds of an atom: 1 for a single bond, 2 for a double bond, and 3 for a triple bond. The valence connectivity indices, ${}^1\chi^v$, developed by Kier and Hall,[30] use the difference in valence electrons and the number of hydrogen atoms to modify the valence parameter for heteroatoms. Finally "edge connectivity" indices were recently tested using bond adjacency rather than vertex adjacency in construction of the modified connectivity indices.[31]

All the above indices, except ${}^1\chi^f$, are based on fixed weights determined by the connectivity of the molecular graph model used. In our view, a better strategy is to introduce weights that make descriptors "flexible", so not only that atoms of different type can adjust their weights in order to yield an optimal characterization of a molecule for a particular property but that they may change values when different properties of the same set of molecules are considered. In general, for a molecule with n different types of atoms, x_1, x_2, ..., x_n, one can have n different weights x_i (i = 1, 2, ..., n); hence, the flexible connectivity index ${}^1\chi^f$ becomes a function of n variables. In the case of sulfides, we consider two variables, the weights of carbon and sulfur atoms. In the case of natural amino acids there are four kinds of atoms: carbon, oxygen, nitrogen, and sulfur; hence, in this case flexible connectivity indices ${}^1\chi^f$ imply optimization of four variables.[24] Even if there are no heteroatoms, variable weights can improve regressions visibly.[25]

It should be noted that while the special types of connectivity indices, viz., ${}^m\chi$, ${}^m\chi^b$, and ${}^m\chi^v$ indices, explore only local regions of the parameter space, the ${}^m\chi^f$ indices are capable of exploring the full potential of the parameter space generated by the presence of heteroatoms in a molecule. The previously mentioned simple connectivity indices and valence connectivity indices can be viewed as a special case of the more general flexible indices ${}^m\chi^f$. Consequently, the flexible indices ${}^m\chi^f$ are expected to be more powerful in predicting molecular properties and biological activities.

Besides the weighted connectivity indices,[21–26] many other topological indices, e.g. the weighted paths $p_t{}^f$,[32–34] the weighted walks, $w_t{}^f$, the weighted Hosoya index Z^f, the weighted Wiener index W^f, and the weighted Balaban index J^f, can be generalized in a similar way.[35] Except for a half-dozen papers of the present authors,[21–26,32–34] use of variable molecular descriptors is in its infancy.

Dramatic improvement in the quality of regressions was obtained by using variable connectivity indices. For example, by introducing a variable parameter x for chlorine in clonidine and clonidine-like imidazolidines (2-(arylimino)-imidazolidines),[21] the value $x = -0.20$ for chlorine produces a regression which, with three weighted connectivity indices, gave better results for the set of clonidine compounds as compared to five descriptors used in a traditional QSAR.[36] This result is particularly striking for this data set, because there are two extreme potency values which would be expected to give much trouble in cross-validation. Use of two variables that differentiate carbon and oxygen in alcohols, with $x = +1.5$ and $y = -0.85$, respectively, reduced the standard error of 7 °C, obtained using the simple connectivity index that does not differentiate carbon and oxygen atoms, to 3.5 °C.[22] In the case of amines, the standard error of 3.48 °C for the boiling point model when ${}^1\chi$ is used has been reduced to 1.91 °C with $x = +1.25$ and $y = -0.65$.[23] The standard error for a quadratic regression using the connectivity index for the boiling points of smaller alkanes is 2.98 °C. When $x = +0.65$ is introduced as a weight, not only is $s = 2.48$ obtained, a reduction by a half-degree Celsius, but higher precision allowed the recognition of an outlier (with an error of over 6 °C), which, when eliminated, further reduced the standard error to an impressive 1.57 °C.[25]

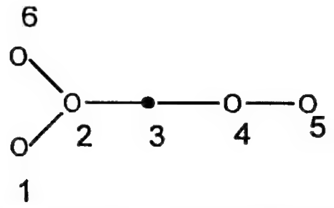


Figure 2. Molecular graph of ethyl isopropyl sulfide and the corresponding numbering of atoms used in Table 2.

Table 2. Adjacency Matrix and Modified Adjacency Matrix for Ethyl Isopropyl Sulfide

adjacency matrix							row sum
	1	2	3	4	5	6	
1	0	1	0	0	0	0	1
2	1	0	1	0	0	0	2
3	0	1	0	1	0	0	2
4	0	0	1	0	1	1	3
5	0	0	0	1	0	0	1
6	0	0	0	1	0	0	1
modified adjacency matrix							**row sum**
	1	2	3	4	5	6	
1	x	1	0	0	0	0	$1 + x$
2	1	x	1	0	0	0	$2 + x$
3	0	1	y	1	0	0	$2 + y$
4	0	0	1	x	1	1	$3 + x$
5	0	0	0	1	x	0	$1 + x$
6	0	0	0	1	0	x	$1 + x$

OPTIMAL DESCRIPTORS FOR SULFUR

We will examine the correlation of the boiling points for sulfides of Figure 1 using functional molecular descriptors and will illustrate the use of a variable connectivity index by considering ethyl isopropyl sulfide (shown in Figure 2 with the numbering of the atoms used). The adjacency matrix and the modified adjacency matrix of ethyl isopropyl sulfide are illustrated in Table 2. If we assume $x = 0$ and $y = 0$, we obtain the usual adjacency matrix of a graph from the row sums of which the simple connectivity index can be directly computed. To obtain the bond contribution for ${}^1\chi$, we use the algorithm $1/(p\ q)^{1/2}$. Here m and n are the respective valences as obtained from the row sums for atoms m and n forming the bond (p, q). When $x \neq 0$ and $y \neq 0$, the corresponding row sums are modified, and instead of the

Table 3. Modified Connectivity Index $^1\chi$ for Ethyl Isopropyl Sulfide with Different Choices of x and y

x	y	$^1\chi\ (x, y)$	x	y	$^1\chi\ (x, y)$
0	−1.00	4.392 51	+0.25	−0.95	2.780 49
0	−1.20	3.297 87	0	0	2.770 06
0	−1.00	3.146 26	+0.25	−0.90	2.753 09
0	−0.95	3.115 31	0	+0.50	2.674 17
0	−0.90	3.086 49	+0.50	−1.00	2.556 25
0	−0.75	3.010 66	+0.50	−0.95	2.528 12
0	−0.50	2.910 56	+1.00	−1.00	2.192 71
0	−0.25	2.832 77	+2.00	−1.00	1.752 29
+0.25	−1.00	2.809 93			

fixed valences p, q, we have the variable valence $(p + x)$, $(q + x)$, or $(q + y)$, depending on the kind of atoms involved. Thus instead of the simple ("fixed") connectivity index $^1\chi = 1/\sqrt{2} + 1/2 + 1/\sqrt{6} + 2/\sqrt{3}$, we have the variable connectivity index given as a function of two variables:

$$^1\chi\ (x, y) = 1/\{(1 + x)(2 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(3 + x)(2 + y)\}^{1/2} + 2/\{(1 + x)(3 + x)\}^{1/2}$$

In Table 3 we listed selected values of the variable $^1\chi$ molecular descriptor for ethyl isopropyl sulfide. As we see, the flexible descriptor is sensitive on the choice of the values for x and y. For a fixed value of x (carbon atom), as y decreases and approaches −1, the magnitudes of the modified connectivity index increase. Similarly for a fixed value of y as x increases the magnitude of the modified connectivity index decreases. An increase and a decrease of the modified index is not so important as is the change of the relative magnitudes of the indices for different molecules.

In Table 4 we have listed the expressions for the modified connectivity indices for the set of $n = 21$ sulfides. In order to illustrate the flexibility of these generalized connectivity indices in Table 5, we listed for the selected values of x and y the numerical values for the variable connectivity indices. Even though for most of the structures the numerical magnitudes have not reversed the relative magnitudes, they altered the magnitudes of the indices for different molecules sufficiently to influence the quality of the regression dramatically. The ratios of the magnitudes of descriptors for different molecules are important for MRA, and these do change. Consider isopropyl propyl sulfide (**14**) and ethyl isobutyl sulfide (**15**) with the boiling points 132.0 and 134.2 °C, respectively. As we can see from Table 5 when $x = -^1/_2$, and $y = -1$, the modified connectivity indices are as follows: 5.059 17 and 5.092 95, giving the quotient 0.9934. However, when $x = +^1/_2$ and $y = -1$ the modified connectivities are as follows: 2.956 25 and 2.992 24, and the quotient decreases to 0.9880. These changes may appear small; however, they are sufficient enough to influence the standard error and make one alternative better than the other. When such changes are summed for all molecules, considerable improvement in the overall standard error is possible.

In Table 6 we show the standard error as a function of the parameters x, y, assuming a quadratic regression using $n = 19$ compounds. We excluded two structures, ethyl butyl sulfide **12** and diisopropyl sulfide **20**, to be discussed later. Using the simple connectivity index, the (0, 0) point in Table 6, the standard error is quite respectable 2.71 °C. Nevertheless this is about twice the magnitude of typical experimental errors reported for boiling points of organic compounds (1−1.5 °C). By keeping x constant and varying y, we see a dramatic reduction of the standard error as we approach the $y = -1$ limit. The standard error for $x = 0$ and $y = -1$ is about 1.5 °C smaller than the initial value ($x = y = 0$). With a further change of both parameters x and y, we find the minimum standard error of 1.326 °C (when $x = +0.25$ and $y = -0.95$). This is less than half of the initial standard error characterizing the "inflexible" connectivity index.

OUTLIERS

Mathematical descriptors, if correctly calculated, are error-free. Hence, if in a correlation between an experimental quantity and mathematical descriptors of one or more points show larger deviation from the regression curve, this can mean two things: Either (1) some experimental data used are in *error* or (2) the descriptors used *fail* to capture some relevant structural feature present in some (and absent in other) molecules.

Whatever is the reason for the departure of a point from the regression line, one can consider such a point as an outlier if the departure from the correlation is more than twice the standard error. In Figure 3 we show the quadratic correlation for sulfides, and in Table 7 we listed the computed boiling point and the residue. As we see from Table 7 ethyl butyl sulfide and diisopropyl sulfide show large departures from the regression. In Table 8 are given the regression equations and the associated statistical parameters for all $n = 21$ sulfides as well as for the cases $n = 19$ sulfides where two outliers have been removed respectively from the set considered.

By eliminating the apparent outliers (**12** and **20**), one substantially reduces the standard error for the quadratic model, as can be seen from the bottom part of Table 8. The standard error for the regression when $n = 19$ reaches the respectable value of 1.33 °C and the correlation coefficient and the Fisher ratio have increased. This signals that the model has improved and that we were justified in eliminating the two outliers.

In Table 9 we listed the optimal connectivity indices for the sulfides considered, the experimental boiling points (BP), the calculated boiling points (BPcalc), the residual of the regression (Res), the cross-validated boiling points (×BPcalc), and the standard error associated with cross-validation (when leaving one entry out). For the two outliers, ethyl butyl sulfide and diisoproyl sulfide, which were excluded when the regression equation was derived, we calculate for the boiling points to be 140.44 and 124.47 °C, respectively. The first of these values is about 4 °C below the reported experimental BP; the second value is almost 4.5 °C higher than the reported experimental BP. The quadratic regression without the data on the two outliers is illustrated in Figure 4.

A closer look at the last column of Table 9, which lists the standard errors associated with the cross-validated regressions, shows (with a single exception **13**, dipropyl sulfide) that the cross-validated standard errors differ about ±0.05 °C from the standard error of the regression (when all $n = 19$ compounds are considered). Hence, disregarding the exception which produced significantly *smaller* standard error, the constancy of the cross-validated standard errors show the robustness of this particular regression.

Table 4. Generalized Flexible Connectivity Indices for $n = 21$ Sulfides (of Figure 1)

1	$2/\{(1 + x)(2 + y)\}^{1/2}$
2	$1/\{(1 + x)(2 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
3	$1/\{(1 + x)(2 + x)\}^{1/2} + 1/(x + 2) + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
4	$1/(x + 2) + 1/\{(2 + x)(2 + y)\}^{1/2}$
5	$2/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(3 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
6	$2/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(3 + x)(2 + y)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + x)\}^{1/2}$
7	$1/\{(1 + x)(2 + x)\}^{1/2} + 2/(2 + x) + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
8	$2/\{(1 + x)(3 + x)\} + 1/\{(3 + x)(2 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
9	$2/\{(1 + x)(2 + x)\} + 2/(2 + x) + 2/\{(2 + x)(2 + y)\}^{1/2}$
10	$3/\{(1 + x)(4 + x)\}^{1/2} + 1/\{(4 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
11	$1/\{(1 + x)(2 + x)\}^{1/2} + 3/(2 + x) + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
12	$2/\{(1 + x)(2 + x)\}^{1/2} + 2/(2 + x) + 2/\{(2 + x)(2 + y)\}^{1/2}$
13	$2/\{(1 + x)(2 + x)\}^{1/2} + 2/(2 + x) + 2/\{(2 + x)(2 + y)\}^{1/2}$
14	$2/\{(1 + x)(3 + x)\}^{1/2} + 1/(2 + x) + 1/\{(3 + x)(2 + y)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
15	$2/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(3 + x)(2 + x)\}^{1/2} + 2/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
16	$2/\{(1 + x)(3 + x)\}^{1/2} + 1/(2 + x) + 1/\{(3 + x)(2 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$
17	$1/\{(1 + x)(2 + x)\}^{1/2} + 1/\{(1 + x)(3 + x)\}^{1/2} + 2/\{(2 + x)(3 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\} + 1/\{(1 + x)(2 + y)\}^{1/2}$
18	$2/\{(1 + x)(2 + x)\}^{1/2} + 1/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(2 + x)(3 + x)\}^{1/2} + 1/\{(3 + x)(2 + y)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2}$
19	$1/\{(1 + x)(2 + x)\}^{1/2} + 3/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(4 + x)(2 + y)\}^{1/2}$
20	$4/\{(1 + x)(3 + x)\}^{1/2} + 2/\{(3 + x)(2 + y)\}^{1/2}$
21	$1/\{(1 + x)(2 + x)\}^{1/2} + 1/\{(1 + x)(3 + x)\}^{1/2} + 1/\{(2 + x)(3 + x)\}^{1/2} + 1/\{(2 + x)(2 + y)\}^{1/2} + 1/\{(1 + x)(2 + y)\}^{1/2}$

Table 5. Modified Connectivity Index $^1\chi$ for Sulfide for a Selection of Choices of x and y

	(0, 0)	(0, −0.5)	(0, −1)	(−0.5, −1)	(+0.5, −1)	(+1, −1)
1	1.414 21	1.632 99	2.000 00	2.828 43	1.632 99	1.414 21
2	1.914 21	2.100 95	2.414 21	3.385 41	1.965 35	1.692 71
3	2.414 21	2.600 95	2.914 21	4.052 08	2.365 35	2.026 04
4	2.414 21	2.568 91	2.828 23	3.942 39	2.297 71	1.971 20
5	2.270 06	2.442 60	2.732 05	3.835 52	2.223 89	1.914 21
6	2.770 06	2.910 56	3.146 26	4.392 51	2.556 25	2.192 71
7	2.914 21	3.100 95	3.414 21	4.718 74	2.765 35	2.359 37
8	2.770 06	2.956 80	3.270 06	4.535 96	2.659 89	2.289 24
9	2.914 21	3.068 91	3.328 43	4.609 06	2.697 71	2.304 53
10	2.560 66	2.724 74	3.000 00	4.216 52	2.442 60	2.103 00
11	3.414 21	3.600 96	3.914 21	5.385 41	3.165 35	2.692 71
12	3.414 21	3.568 91	3.828 43	5.275 37	3.097 71	2.637 86
13	3.414 21	3.568 91	3.828 43	5.275 37	3.097 71	2.637 86
14	3.270 06	3.410 56	3.646 26	5.059 17	2.956 25	2.526 04
15	3.270 06	3.424 76	3.684 27	5.092 95	2.992 24	2.558 73
16	3.270 06	3.456 80	3.770 06	5.202 63	3.059 89	2.613 57
17	3.308 06	3.494 80	3.808 06	5.312 63	3.077 91	2.623 61
18	3.308 06	3.448 57	3.684 27	5.169 18	2.974 27	2.536 08
19	3.060 67	3.192 71	3.414 21	4.773 51	2.774 96	2.381 50
20	3.125 90	3.252 21	3.464 10	4.842 62	2.814 79	2.414 21
21	3.346 07	3.518 61	3.808 06	5.388 87	3.059 94	2.600 95

Table 6. Standard Error of the Regression for Different Choices of the Variable Parameters x and y

	−0.5	0	+0.25	+0.50	+1	+2
+0.50		3.273				
0		2.711				
−0.25		2.363				
−0.50		1.966				
−0.75		1.558				
−0.90		**1.382**	1.347			
−0.95		1.356	1.326	1.380		
−1	2.256	1.357	1.327	1.327	1.570	2.042
−1.2		1.720				

We believe that it may be possible to further improve the regression. A close inspection of residuals shows, with very few exceptions, that all linear structures have positive residual, while all branched structures show a negative residual. This suggests the possibility for further reduction of the standard error (particularly if the exceptions are viewed as outliers). However, such refinements should be attempted when a larger set of compounds is considered in order to see if the observed trend is genuine or not.

Finally, as a warning, we should add that when using flexible descriptors, elimination of outliers may influence the optimal values for the parameters x, y, though not necessarily dramatically.

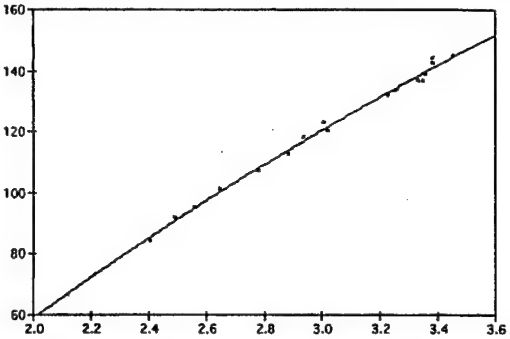


Figure 3. 3. Quadratic regression for the boiling points of $n = 21$ sulfides against the optimal connectivity index ($x = +0.25$, $y = -0.95$).

Table 7. Calculated Boiling Points (BPcalc) and the Residual of the Regression (Res), When All $n = 21$ Sulfides Are Considered

	BP	BPcalc	Res
1	37.3	38.44	−1.14
2	66.6	65.53	+1.07
3	95.5	94.86	+0.64
4	92.0	90.42	+1.58
5	84.4	84.81	−0.41
6	107.4	108.01	−0.61
7	123.2	121.09	+2.11
8	112.5	114.14	−1.64
9	118.5	117.14	+1.36
10	101.5	100.04	+1.46
11	145.0	144.21	+0.79
12	144.2	140.75	**+3.45**
13	142.8	140.75	+2.05
14	132.0	132.73	−0.73
15	134.2	134.52	−0.32
16	137.0	138.12	−1.12
17	139.0	139.40	−0.40
18	133.6	134.05	−0.45
19	120.4	121.82	−1.42
20	120.0	124.31	**−4.31**
21	137.0	138.94	−1.94

CONCLUDING REMARKS

Several criticisms could be raised concerning the outlined work:[37] Is it appropriate to refer to MRA using flexible

Table 8. Linear and Quadratic Regressions for Sulfides[a]

n	model	coeff x	coeff x^2	constant	r	s	F
21	linear	60.1981		−61.3339	0.9959	2.61	2291
21	quadratic	102.8180	−7.8615	−117.0919	0.9981	1.83	2328
21	orthogonal	60.1981	−7.8615	−61.3339	0.9981	1.83	2328
19	linear	60.1057		−60.9916	0.9961	2.59	2180
19	quadratic	108.9647	−9.0423	−124.6847	0.9990	1.33	4192
19	orthogonal	60.1057	−9.0423	−60.9916	0.9990	1.33	4192

[a] The top part gives the regression equations and the statistical parameters for all $n = 21$ sulfides; the bottom part gives the equations when two outliers are excluded.

Table 9. Optimal Connectivity Indices for the Sulfides Considered, the Experimental Boiling Points (BP), the Calculated Boiling Points (BPcalc), the Residual of the Regression (Res), the Cross-Validated Boiling Points (×BPcalc), and the Standard Error of Cross-Validated Boiling Points

	(+0.25, −0.095)	BP	BPcalc	Res	×BPcalc	xstd error
1	1.745 75	37.3	37.98	−0.68	40.65	1.31
2	2.119 76	66.6	65.66	+0.94	65.41	1.34
3	2.564 20	95.5	95.27	+0.23	95.23	1.37
4	2.493 37	92.0	90.82	+1.18	90.60	1.33
5	2.406 48	84.4	85.17	−0.77	85.30	1.35
6	2.780 49	107.4	108.38	−0.98	108.51	1.34
7	3.008 65	123.2	121.30	+1.90	120.86	1.27
8	2.885 55	112.5	114.45	−1.95	114.66	1.26
9	2.938 21	118.5	117.41	+1.07	117.31	1.34
10	2.647 84	101.5	100.44	+1.06	100.28	1.38
11	3.453 09	145.0	143.76	+1.24	143.42	1.32
12	3.382 66	144.2				
13	3.382 66	142.8	140.44	+2.36	138.86	1.20
14	3.224 94	132.0	132.68	−0.68	132.74	1.38
15	3.259 56	134.2	134.42	−0.22	134.44	1.37
16	3.329 99	137.0	137.90	−0.90	138.01	1.35
17	3.355 50	139.0	139.14	−0.14	139.16	1.37
18	3.250 44	133.6	133.96	−0.36	134.00	1.37
19	3.021 85	120.4	122.02	−1.62	122.16	1.30
20	3.067 22	120.0				
21	3.346 37	137.0	138.69	−1.69	138.94	1.29

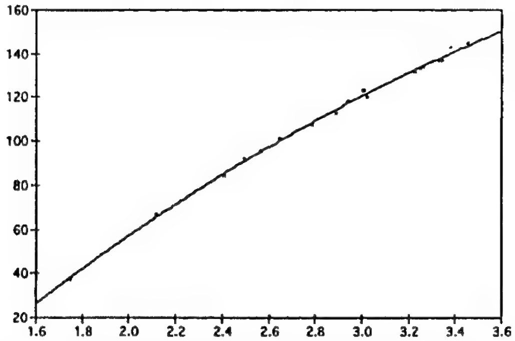


Figure 4. 4. Quadratic regression for the boiling points of $n = 19$ sulfides against the optimal connectivity index ($x = +0.25$, $y = -0.95$). Outliers excluded **12** and **20**.

descriptors as "high-quality regression", or should it be called "high specialty SAR"? Is one justified to arrive at low standard error by "trimming the data set and by tweaking the descriptor"? Would the model be any good to predict boiling points even for other sulfides? Is the approach general enough and sufficiently justified if we were to use QSAR models for real world problems? Why not consider more extensive study on a larger set of data to strengthen the case? What is the use of a model developed by considering a quite small, homogeneous set of compounds? Is developing a fit with standard error less than that of the experimental error (if that can be achieved) overfitting?

We respond to these question one by one. Variable connectivity indices (and related variable indices) constitute a general class of descriptors as compared to the special class of descriptors used in QSAR (e.g. indicator variables used in some QSAR, or hydrogen bonding descriptors used in CODESSA) for which the attribute "high specialty" holds. Concerning the problem of identifying outliers, these are well-defined as points that are beyond 2 standard deviations. There are no good reasons for their inclusion in the data set, despite that their departure from the regression need not be due to experimental error. Most often they are not. The occurrence of outliers may be a signal that the set of descriptors used to characterize molecules failed to characterize some special structural features which are important for outliers but not for most of other molecules in the set. A close look at outliers may help one to recognize such features, if they are not obvious. For example, correlation of the boiling points of smaller alkanes[25] shows only 2,2,3,3-tetramethylbutane was identified as an outlier (with deviation of over 6 °C), while the standard error was 2.48 °C. By removing this compound, standard error dropped to 2 °C. Hence, a single compound in a set of 20 was able to increase the standard error almost by $^1/_2$ °C. Why should this compound that has *additional* structural features (significant overcrowding of methyl groups and a quaternary CC bond) absent in the rest be included if one is interested in predicting the boiling point of a compound which has no overcrowded methyl groups and no quaternary CC bond?

Smaller sulfides considered (and the same has been the case with smaller alkanes or amino acids) are molecules of similar size. To consider large selection of compounds necessarily brings the dominant role of molecular size into focus as important feature. Before we do this, we should investigate to what extent the variable weights may depend on the size of the molecule. At the moment this is an unresolved problem, which is the main reason for restricting attention to smaller sets of compounds with similar size. We should add that it is not uncommon in QSAR to consider smaller sets of compounds, often because of limited data. For example in a recent review of comparative QSAR Hansch and co-workers[38,39] gave results for 189 regressions in which only 33 had more than 20 compounds in the set, and 156 had less than 20 compounds, that is, less than the number of sulfides considered in this paper. If compounds are well-selected, the resulting regressions may be of interest. We gave here the results for *smaller* sulfides. If one is interested in larger sulfides, one should select those, and if one is interested in all sulfides, one should combine them all. But again a question can be raised: If one is interested in predicting the boiling point of *smaller* sulfides, why does one need information of compounds that are *twice* its size? It is a matter of philosophy, and while we appreciate the merits of studying a large data basis, we also appreciate the advantages of studying small homogeneous sets of compounds. Such a study focuses attention at different aspects of structural chemistry. In fact, one of the present author made numerous studies on the large set of compounds using diverse types of molecular descriptors.[40−45]

Concerning "overfitting", which is clearly undesirable, we would like to point out that this is out of the question when one uses a single descriptor. Overfitting is a danger in multiple regression analysis when one uses too many descriptors and has too few data. One cannot have overfitting with a single descriptor. This problem received some attention.[46] Does the variation of descriptors during the regression poses such a threat? Definitely so, just as a selection of descriptors from a large pool of descriptors (e.g. in CODESSA software) does the same. The difference between the two is that typically when using variable connectivity index, one generates about 40 different numerical alternative descriptors to choose from, CODESSA typically chooses a half-dozen descriptors from a pool of some 400 descriptors!

Finally we have to emphasize that while the idea of modifying chemical graph descriptors to differentiate heteroatoms is not new, as is well-illustrated by the pioneering work of Kier and Hall on valence connectivity indices,[28] the idea of modifying chemical graph descriptors to differentiate heteroatoms during the search for the best regression; that is, the idea of variable topological indices, is new.

ACKNOWLEDGMENT

This paper is contribution number 265 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grant F49620-94-1-0401.

REFERENCES AND NOTES

(1) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; Chapter 10, pp 225−273.

(2) Balaban, A. T. Historical developments of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devilers, J., Balaban, A. T., Eds., in press.

(3) Randić, M. Topological Indices. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; pp 3018−3032.

(4) Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devilers, J., Balaban, A. T., Eds.

(5) Randić, M.; Novic, M.; Vracko, M. *Molecular Descriptors, New and Old*; Lecture Notes in Chemistry; Springer: Berlin, submitted for publication.

(6) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.

(7) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−370.

(8) Randić, M. Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14*, 363−370.

(9) Randić, M. Curve fitting paradox. *Int. J. Quantum Chem, Quantum Biol. Symp.* **1994**, *21*, 215−225.

(10) Amić, D.; Davidović-Amić, D.; Jurić, A.; Lučić, B.; Trinajstić, N. Structure−activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.

(11) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. The structure−property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.

(12) Šoškic, M.; Plavšic, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.

(13) Katritzky, A. R.; Lobanov, V.; Karelson, M. *CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis)*; University of Florida: Gainesville, FL, 1994.

(14) Basak, S. C. *POLLY*; (Natural Resources Research Institute, Duluth, University of Minnesota: Duluth, MN, 1988.

(15) Hall, L. H. *MOLCONN-X*; Hall Associates Consulting, Quincy: MA, 1991.

(16) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(17) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(18) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(19) Randić, M. Linear combinations of path numbers as molecular descriptors. *New J. Chem.* **1997**, *21*, 945−951.

(20) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations between chemical structure and normal boiling points of acyclyc ethers peroxides, acetals, and their sulfur analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237−244.

(21) Randić, M. Novel graph theoretical approach to heteroatoms in QSAR. *Chemom. Intel. Lab. Syst.* **1991**, *10*, 213−227.

(22) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1991**, *12*, 970−980.

(23) Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quantum Chem.* **1998**, *70*, 1209−1215.

(24) Randić, M.; Mills, D.; Basak, S. C.; Pogliani, L. On characterization of physical properties of amino acids. *New J. Chem.*, submitted for publication.

(25) Randić, M. High quality structure-property regressions. Boiling points of smaller alkanes. *New J. Chem.*, in press.

(26) Randić, M.; Basak, S. C.; Pompe, M.; Novic, M. Prediction of Gas Chromatographic Retention Indices Using Variable Connectivity Index. *Acta Chim. Slov.*, submitted for publication.

(27) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. Molecular Connectivity I. Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971−1974.

(28) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity Series Applied to Density. *J. Pharm. Sci.* **1975**, *65*, 1226−1230.

(29) Basak, S. C.; Magnuson, V. R. Determining structural similarity of chemicals using graph-theoretical indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.

(30) Kier, L. B.; Hall. H. L. Molecular Connectivity VII. Specific Treatment of Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.

(31) Estrada, E. Edge adjacency relationships and novel topological index related to molecular volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31−33.

(32) Randić, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261−266.

(33) Randić, M.; Basak, S. C. Multiple regression analysis with optimal molecular descriptors. *SAR QSAR Environ. Res.*, in press.

(34) Randić, M.; Pompe, M. On characterization of CC double bond in Alkenes. *SAR QSAR Environ. Res.* **1999**, *10*, 451−471.

(35) Randić, M.; Pompe, M. Work in progress.

(36) Timmermans, B. M. W. M.; van Zweiten, P. A. Quantitative structure−activity relationship in centrally acting imdazolidines structurally related to clonidine. *J. Med. Chem.* **1977**, *20*, 1636−1644.

(37) Based on the reviewers' comments.

(38) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology;* American Chemical Society: Washington D.C., 1995.

(39) Kubinyi, H. *Hansch Analysis and Related Approaches;* VCH: Weinheim, Germany, 1993.

(40) Basak, S. C.; Niemi, G. J.; Veith, G. D. Optimal characterization of structure for prediction of properties. *J. Math. Chem.* **1990**, *4*, 185−205.

(41) Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. Prediction of octanol/water partition coefficient (K_{OW}) with algorithmically derived variables. *Environ. Toxicol. Chem.* **1991**, *10*, 893−900.

(42) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: Hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.

(43) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph-theoretic and geometrical molecular descriptors in structure−activity relationships. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York 1997.

(44) Basak, S. C.; Niemi, G. J.; Veith, G. D. Recent developments in the characterization of chemical structure using graph-theoretical indices. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 235−277.

(45) Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of complement−inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted for publication.

(46) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238−1244.

CI990115Q

APPENDIX 1.8 Multiple regression analysis with optimal molecular descriptors

SAR and QSAR in Environmental Research
2000, Vol. 11, pp. 1–23
Reprints available directly from the publisher
Photocopying permitted by license only

MULTIPLE REGRESSION ANALYSIS WITH OPTIMAL MOLECULAR DESCRIPTORS

M. RANDIC[a,*] and S. C. BASAK[b]

[a]*Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311 (USA) and National Institute of Chemistry, 1001 Ljubljana, POB 3430 (Slovenia);* [b]*Natural Resources Research Institute, The University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811 (USA)*

(Received 17 September 1998; In final form 28 August 1999)

We consider construction of optimal molecular descriptors to be used for multiple regression analysis of several properties of alcohols. The descriptors are obtained by considering shorter paths with variable weight x for carbon-oxygen bond in alcohol. In particular we consider as molecular descriptors paths of length 1, 2 and 3. The multiple regression analysis of the following molecular properties was examined: $-\log\ S$ (S = solubility), CSA (cavity surface area), $\log\ P$ (P = octanol/water partition), and $\log\gamma$ (γ = infinite solution activity coefficient). By minimizing the standard error of the regression for each property we found optimal variable weight.

Keywords: Variable molecular descriptors; weighted paths; MRA; orthogonal descriptors; alcohol properties

INTRODUCTION

Study of structure-property and structure-activity relationship continues to attract considerable attention in chemical literature. Various statistical methods have been found useful in such studies, including the Principal Component Analysis (PCA) [1], the Pattern Recognition (PR) [2], the Partial Least Square method (PLS) [3], the Artificial Neural Networks (ANN) [4]. The oldest data reduction method, the Multiple Regression Analysis (MRA) [5], continues to be widely used. Most applications of MRA to QSAR and SAR can be classified into one of two types:

[*] Corresponding author.

(I) Examination of large number of diverse and heterogeneous structures;
(II) Study of smaller number of homogenous structures.

Each of these studies have their merits and will continue to be pursued. In both cases often one starts screening a large pool of molecular descriptors from which one selected smaller number of descriptors that are used for construction of regression equations, or construction of principal components. An alternative, particularly suitable when one study smaller number of structurally related compounds, is to focus attention on only few molecular descriptors which are general enough to be used in different applications [6, 7]. Such descriptors were referred to as basis descriptors in analogy with basis vectors in linear algebra. Advantage of basis descriptors is that they facilitate comparative analysis, because the same descriptors are used in different applications, for different molecules and different properties. For example, Kier and Hall [8] used different combinations of the connectivity indices for the best correlation of alkane heats of atomization and alkane heats of formation. If, however, one restrict search for best correlation for the two properties to the same connectivity indices one finds that the two properties are strictly collinear, the fact that is obscured when one uses different descriptors because the two samples of structures are somewhat different.

Despite its wide use MRA was viewed by some as deficient, because as a rule introduction of an additional descriptor in the analysis causes dramatic changes of the contributions of already used descriptors. Because of this pronounced instability of the regression equations it is not possible to interpret the results in terms of the relative role of the descriptors used. This deficiency (which incidentally is not confined solely to MRA) has been traced to mutual interrelation of descriptors [9–13]. If the descriptors used are to a greater extend independent of one another one observes but a minor variations of the coefficients of the regression equation if a descriptor is included or excluded. However use of moderately and highly intercorrelated descriptors, which often cannot be avoided, results in pronounced instability of the regression equation. This is particularly visible when one introduces descriptors one at a time in a stepwise regression.

This very unsatisfactory affair has been tolerated because despite the instability of the regression equations each additional relevant descriptor decreases the standard error of prediction for the property considered. Thus the equation offers useful predictions but it does not offer useful interpretation. This MRA nightmare — as some have referred to it — is no more. With introduction of orthogonalization procedure for molecular descriptors not only that the regression equation becomes stable but the error of the coefficients reduces with introduction of each additional relevant descriptor [12]. While some have recognized the significance of using orthogonal molecular descriptors [14–16] apparently others still do not appreciate or are unaware of the novel situation, which for the first time makes possible to interpret the relative contributions of descriptors used.

We will refer to MRA using molecular descriptors as MORA, the Multivariate Orthogonal Regression Analysis. It has been shown that MORA and MRA remain related so that one can obtain orthogonalized regression equation form MRA by stepwise regression [9, 10]. With this remedy MRA not only remains a very viable data reduction method for QSAR and QSPR, but in some way may again become the method of choice, despite the fact that researchers in the field are free to be reluctant to use a new method! In our opinion MORA has an important advantage over PCA. MORA, just as PCA, uses orthogonal descriptors but in contrast to PCA the descriptors used in MORA can be interpreted in terms of the structural meaning of the initial descriptors. In contrast the linear combinations that define the principal components have, at best, a vague interpretation (i.e., as bulk, cohesiveness, etc.). Not only that it is hard to visualize what such linear combinations of descriptors represent, the descriptors that define the principal components are themselves not orthogonal, despite that the principal components are mutually orthogonal. So we are in no better situation, as far as an interpretation of the results of PCA is concerned, then we have been with MRA in the time of instabilities of the regression equations!

OPTIMAL MOLECULAR DESCRIPTORS

With hundreds of molecular descriptors available [17–19] immediately one is confronted with decision concerning selection of descriptors. The choices to consider are: (a) select a subset of "the best" descriptors from a large pool of available descriptors; (b) use a limited set (of "well ordered" structurally related descriptors, the basis; (c) use as few as possible descriptors that are suitably optimized for the particular application. We will refer to the last alternative as use of optimal molecular descriptors. In the first case we put "the best" under quotes because the outcome will depend on the criteria used to select descriptors. Current practice that many adopted of excluding descriptors that are highly intercorrelated to descriptors already selected, as argued elsewhere [20, 21], has no theoretical justification. We also put "well ordered" under quotes because ordering of descriptors will influence interpretation, even though it will not influence the statistical parameters of the regression analysis.

Optimization of molecular descriptors is relatively novel technique in QSAR and SAR that has been for the most part overlooked. It is generally recognized that the presence of heteroatoms in a molecule requires use of additional molecular descriptors. However, these additional descriptors to be used for C–X bonds (X can be O, N, Cl, etc.) are usually in advance prescribed, using some physicochemical analogy or data. For example, Kier and Hall introduced the valence connectivity indices by assigning to atoms valence parameter based on the count of valence electrons of each atom [22]. Another possibility, perhaps not so widely known, uses covalent radii of carbon and other atoms in deriving parameters to differentiate atoms of different kind [23]. In contrast one of the present authors considered variable weight as an entry on the main diagonal of the adjacency matrix of a molecular graph. For example, for ethyl alcohol one would have for so generalized adjacency matrix:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & y \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} x & 1 & 0 \\ 1 & x & 0 \\ 0 & 1 & y \end{pmatrix}$$

Here x, and y represent variables describing carbon and oxygen atom respectively. Using x and y as variables one can construct the connectivity indices (or connectivity weighted paths) and search for best values of x and y that would minimize the standard error in the regression analysis of the property of interest [24]. For example, in the case of boiling points of alcohols one finds $x = 1.50$ and $y = -0.85$ to result in the smallest standard error. Use of the diagonal entries has been already considered some time ago in chemical documentation by Spialter who developed alphanumeric matrices for a representation of chemical structure [25]. The difference is however, that rather than using symbols C and O (corresponding to x and y) here we search for numerical parameters that result in the best regression. In the case of chlorine atom the diagonal entry $y = -20$ [26] was found to give a better regression that approaches based on the "traditional" (i.e., the approaches following Hansch's methodology [27]) molecular descriptors. Similarly, in the case of nitrogen containing molecules the diagonal entries $x = 1.25$ for carbon and $y = -0.65$ for nitrogen give the optimal solution for the boiling points of amines [28].

All the above cases relate to the connectivity indices and paths when weighted using the same weighting algorithm. However, variable descriptors can be constructed for other topological indices besides the connectivity indices. Construction of these variable generalizations of the Wiener index [29] and the Hosoya index [30] have been recently outlined [31]. Recently variable weights have been considered for path numbers [32, 33]. We continue with exploration of optimally weighted path numbers for characterization of molecules in this article.

WEIGHTED PATH NUMBERS

Path numbers have been suggested fifty years ago by Platt as potentially useful molecular descriptors [34]. Apparently the contribution of Platt, despite its importance, has been overlooked till a revived interest in chemical graph theory emerged in mid 1970's. Apparently through a series of papers [35–43] Randić and Wilkins resurrected path numbers and have illustrated use of paths for characterization of molecules and their fragments. Later Randic and coworkers [44–49] introduced weights for paths of different length by weighting the contributions of bonds and longer paths by using $1/\sqrt{(m\ n)}$ as the weight for individual bonds involved. Weighted paths are also implied in construction of higher order connectivity indices [50]. All these cases, however, used rigidly prescribed weighting scheme, which, once adopted does not change.

As already mentioned the use of the diagonal entries of the adjacency matrix as variable input initiated construction of new kind of molecular descriptors. In contrast to hitherto used topological indices and other descriptors the new descriptors have an inherent flexibility that allows them to be constructed so to minimize the standard error in a regression. Very recently this kind of flexibility associated with variable weights has been extended to construction of weighted molecular paths. This has lead to generalized Wiener number [32], and generalized path numbers [33] already mentioned. Formally the Wiener number can be written as:

$$W = 1\ p_1 + 2\ p_2 + 3\ p_3 + 4\ p_4 + \cdots + k\ p_k$$

where $p_1, p_2, p_3, \ldots$ are the number of paths of length one, length two, length three, etc. The above can be viewed as dot product of vectors $L = (1, 2, 3, 4, \ldots k)$ and vector $P = (p_1, p_2, p_3, \ldots p_k)$. If now one introduces vectors L^m of the form $(1^m, 2^m, 3^m, \ldots k^m)$ the dot product W becomes function of the exponent m, i.e., instead of W we have now W(m). Here one treats m as variable and, for example, in the case of alkanes the best quadratic fit of motor octane numbers is obtained when $m = -1.50$ while the best quadratic fit for the boiling points of alkanes is obtained when $m = 1.90$.

Randić and Pompe [33] considered a different kind of weights for paths when examining the molar refraction of unsaturated hydrocarbons. They associated the weight x to individual C=C bond in alkenes and assigned the weight x to all paths that involve C=C bond. This approach applies equally to characterization of heterobonds, as illustrated by Randic and Basak when revisiting the correlation of the boiling points of alcohols [51]. In Table I we give the enumeration of weighted path for 3-methyl-1-butanol and 2-pentanol, which if one does not differentiates CC and CO bonds would give the same path count 5, 5, 3, 2, instead of $4 + x$, $4 + x$, $2 + x$, $2x$ and $4 + x$, $3 + 2x$, $2 + x$, $1 + x$ respectively.

TABLE I Weighted paths for 3-methyl-1-butanol and 2-pentanol

3-Methyl-1-butanol					*2-Pentanol*				
atom	p_1	p_2	p_3	p_4	atom	p_1	p_2	p_3	p_4
1	1 + x	1	2		1	1	1 + x	1	1
2	2	2 + x			2	2 + x	1	1	
3	3	1	x		3	2	2 + x		
4	1	2	1	x	4	2	1	1 + x	
5	1	2	1	x	5	1	1	1	1 + x
6	x	x	x	2x	6	x	2x	x	x
Molecule:									
	4 + x	4 + x	2 + x	2x		4 + x	4 + 2x	2 + x	1 + x

Clearly when x = 1 the two path vectors are identical, but already setting x = 1.1 or x = 0.9 results in differentiation between the two isomers. In the case of molar refraction of heptene isomers when using three path numbers the value of x = 0.6 leads to an impressive reduction in the standard error (s = 0.08).

REVIEW OF THE EXPERIMENTAL DATA USED

QSAR and SAR studies often point to FEW experimental points that do not fit well the derived correlation. So identified outliers are then omitted from correlations with some justification, even though the source for the disagreement is not known and need not be attributed to presumed experimental error. It is possible that some outliers have unrecognized structural features which the descriptors used can not adequately characterize that makes them exceptional. Nevertheless, by being different than other compounds under analysis, the outliers may legitimately be eliminated from considerations. In our study, as will be seen shortly we were able to identify one such outlier even before starting the regression analysis. Having several properties of alcohols available we decided first to review property-property correlations of alcohols to be studied. This pointed to a discrepancy for the experimental data of 2-hexanol.

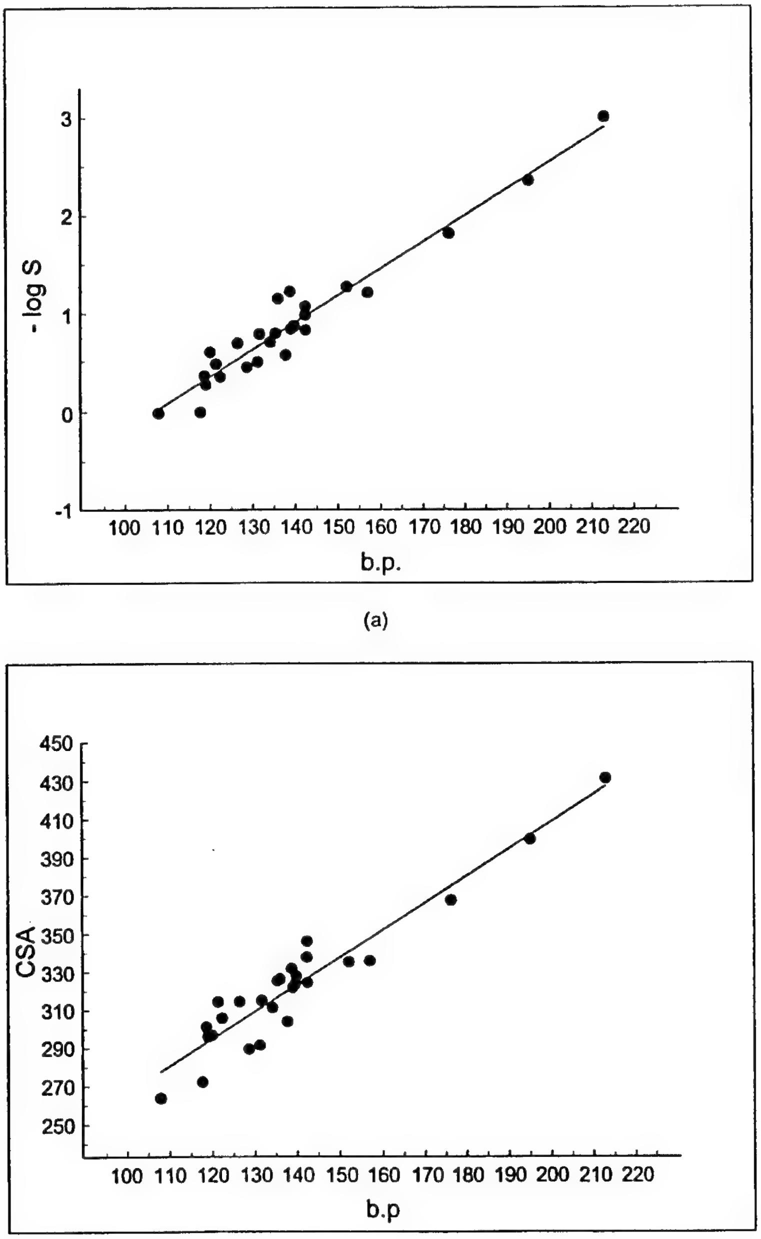
We have selected the following properties of alcohols: (a) water solubility ($-\log S$); (b) cavity surface area (CSA); (c) octanol water partition ($\log P$); and (d) infinite dilution activity coefficient ($\ln\gamma$). Already in ref. [51] we examined the boiling points of alcohols. All these properties have been recently studied by MRA using alternative molecular descriptors by Cao and Li for $-\log S$, CSA, and $\log P$ [52], and by Mitchell and Jurs for $\ln\gamma$ [53]. A set of n = 50 alcohols were used when considering $-\log S$ and CSA, a set of n = 38 alcohols were used in $\log P$ study and a set of n = 43 alcohols were used for $\ln\gamma$ study. In Table II we collected the experimental data for a subset of alcohols studies in ref. [51–53].

TABLE II Common experimental data for different sets of alcohols studied (including boiling points studied in ref. [20])

Alcohol	*−log S*	*CSA*	*log P*	*ln γ*	*b. p.*
1-butanol	0	272.1	0.88	3.92	117.7
2M-1-propanol	−0.01	263.8	0.61	3.89	107.9
1-Pentanol	0.58	303.9	1.40	5.29	137.8
3M-1-butanol	0.51	291.4	1.14	5.34	131.2
2M-1-butanol	0.46	289.4	1.14	5.08	128.7
2-Pentanol	0.28	295.9	1.14	4.57	119.0
1-Hexanol	1.21	335.7	2.03	6.68	157.0
2-Hexanol	0.87	327.7	1.61	5.64	139.9
3-Hexanol	0.80	325.3	1.61	5.85	135.4
3M-3-pentanol	0.36	305.8	1.39	4.85	122.4
2M-2-pentanol	0.49	314.3	1.39	5.14	121.4
2M-3-pentanol	0.70	314.3	1.41	5.63	126.5
3M-2-pentanol	0.71	311.3	1.41	5.66	134.2
2,3MM-2-butanol	0.37	301.2	1.17	4.88	118.6
3,3MM-2-butanol	0.61	296.7	1.19	5.43	120.0
4M-2-pentanol	0.79	314.9	1.41	5.86	131.7
1-Heptanol	1.81	367.5	2.34	8.09	176.3
2M-2-hexanol	1.07	346.1	1.87	6.49	142.5
3M-3-hexanol	0.98	337.7	1.87	6.29	142.4
3E-3-pentanol	0.83	324.4	1.87	5.94	142.5
2,3MM-2-pentanol	0.87	323.8	1.67	6.02	139.7
2,3MM-3-pentanol	0.84	321.8	1.67	5.96	139.0
2,4MM-3-pentanol	1.22	331.7	1.71	6.82	138.8
2,2-MM-3-pentanol	1.15	326.1	1.69	6.66	136.0
1-Octanol	2.35	399.4	2.84	9.56	195.2
2,2,3MMM-3-pentanol	1.27	335.2	1.99	6.95	152.2
1-Nonanol	3.00	431.2	3.15	11.0	213.1

In Figure 1 we illustrate the correlations for the properties listed in Table II. In Figure 1a — Figure 1d we show correlation of the four properties considered here ($-\log$ S, CSA, log P and $\ln\gamma$) with the boiling points of alcohols. The correlations between the four properties among themselves (included in Table III) show similar behavior, similar scatter of points, with a single exception. The exceptional is the correlation between the two solubilities $-\log$ S and $\ln\gamma$, shown in Figure 1e, which display extremely high correlation. While for most other property-property correlations of Table III the regression coefficients is between $r = 0.950$ and $r = 0.990$ the correlation of $-\log$ S and $\ln\gamma$ have $r = 0.998$. That $-\log$ S and $\ln\gamma$ make exceptional correlation is even better reflected in Fisher ratio, which for all mutual property-property correlations is below 500, but $-\log$ S and $\ln\gamma$ have impressive F close to 7000.

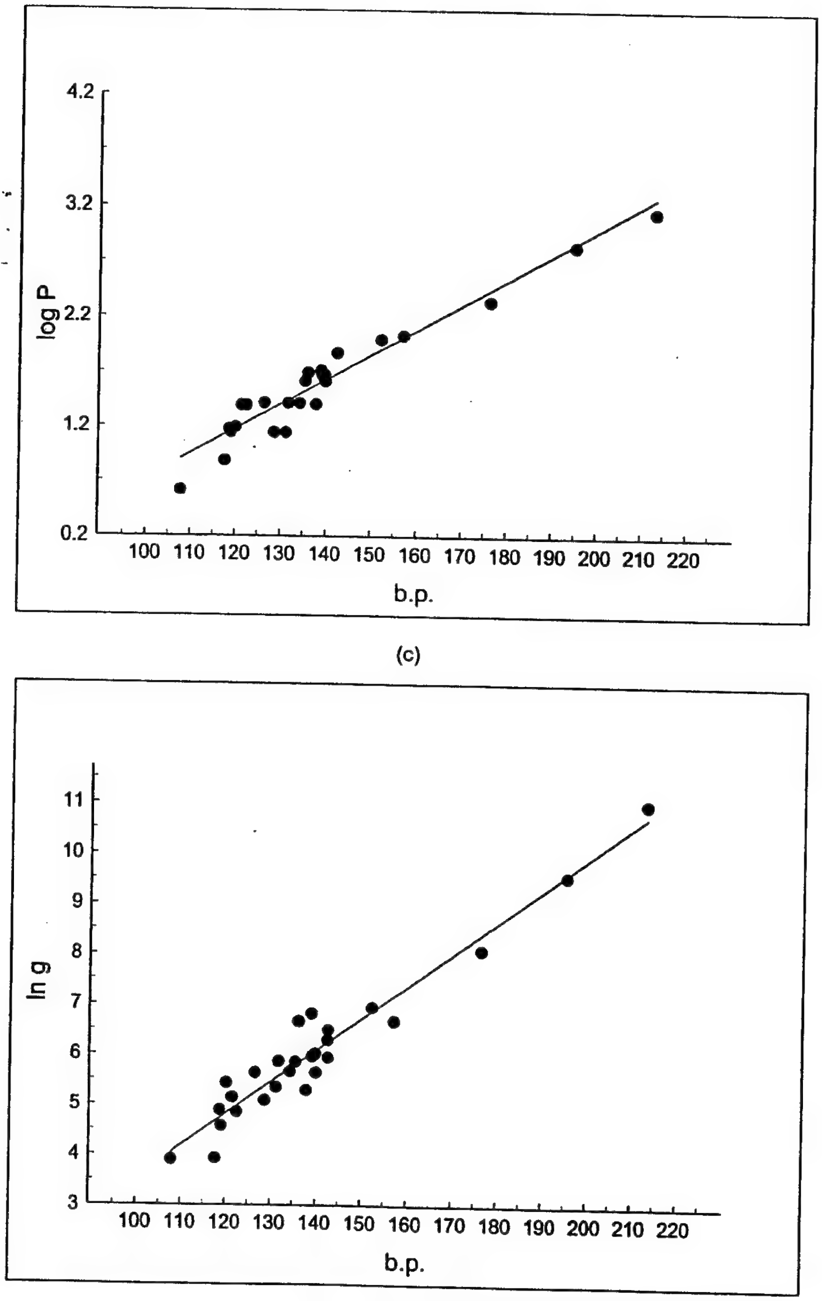
It is clear from Figure 1e that a single point appears to be an outlier, most likely an experimental error either in $-\log$ S or $\ln\gamma$. When this point (that belongs to 2-hexanol) is eliminated the revised regression (shown in the lower part of Table III and indicated by an asterisk) of $-\log$ S and $\ln\gamma$ shows a



(a)

(b)

FIGURE 1 Correlations between different experimental properties of smaller alcohols. Illustrations (a) — (d): Correlations with their experimental boiling points: Negative logarithm of solubility S; critical surface area CSA; logarithm of octanol/water partition P; natural logarithm of solublility γ, respectively. Illustration (e): Correlation between the solubilities $-\log$ S and $\ln \gamma$.



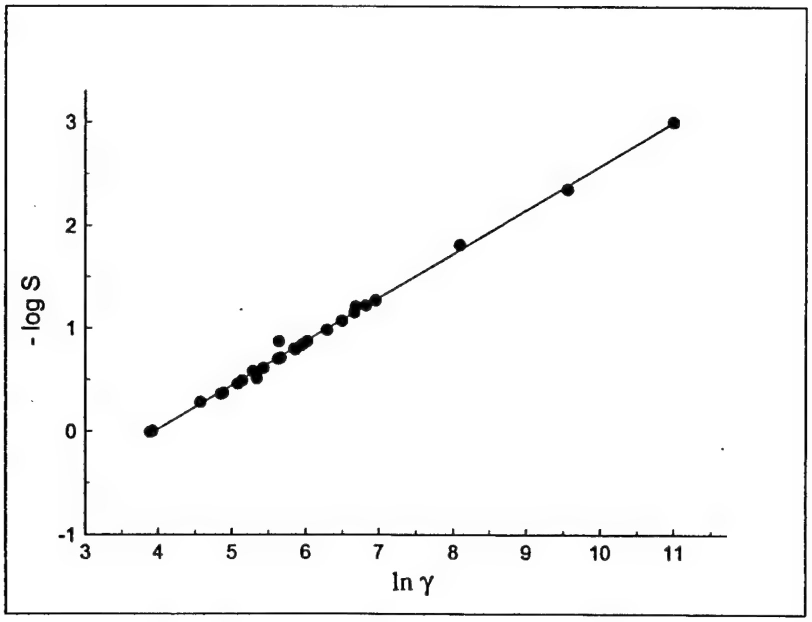
(c)

(d)

FIGURE 1 (*Continued*).

TABLE III Comparison of correlation parameters for property-property correlations of alcohols. The asterisk (*) indicates the regression in which outlier was removed

Property-property	*r*	*s*	*F*
$-\log$ S/b.p.	0.9705	0.161	404
CSA/b.p.	0.9499	11.15	231
log P/b.p.	0.9620	0.153	310
ln g/b.p.	0.9669	0.400	359
$-\log$ S/ $\ln\gamma$	0.9982	0.040	6873
CSA/$\ln\gamma$	0.9721	8.372	429
log P/ $\ln\gamma$	0.9645	0.147	334
$-\log$ S/ log P	0.9674	0.169	364
CSA/ log P	0.9843	6.296	778
$-\log$ S/CSA	0.9752	0.1479	486
$-\log$ S*/ $\ln\gamma$*	0.9993	0.026	16,752



(e)

FIGURE 1 (*Continued*).

dramatic improvement (r = 0.999 and F is over 16,750). This further supports the suspicion that one of the experimental results for 2-hexanol was in error.

That the selected alcohol properties show limited correlation (except for already mentioned intercorrelation of the two solubilities) points to the fact that different properties are dominated by different structural factors and will require different molecular descriptors. Clearly the considered properties can not be reduced to the same structural features, which for itself speaks why we need different molecular indices and should continue to design novel topological descriptors.

That 2-hexanol is an outlier is even better visible in Figure 2 in which we show the same regressions between $-\log$ S and $\ln\gamma$ but have limited the set of alcohols to isomers of 1-hexanol. In this way we eliminated the dominant role of molecular size (since we consider only alcohols having the same number of carbon atoms). In Table IV we give the statistical data for regressions the corresponding regressions when considering n = 10 hexanols. As we see from Table IV the statistical parameters have changed dramatically not only because we have a smaller sample but it is much harder to fit data for molecules of a same size than correlating data for molecules of different size. The standard error which now reflects the isomeric variations has decreased but the correlation coefficient also decreased, because it is more difficult to correlate that part of a property that does not depend on size than the part of the property that is size dependent. That 2-hexanol is outlier is now reflected in an unusual increase (by

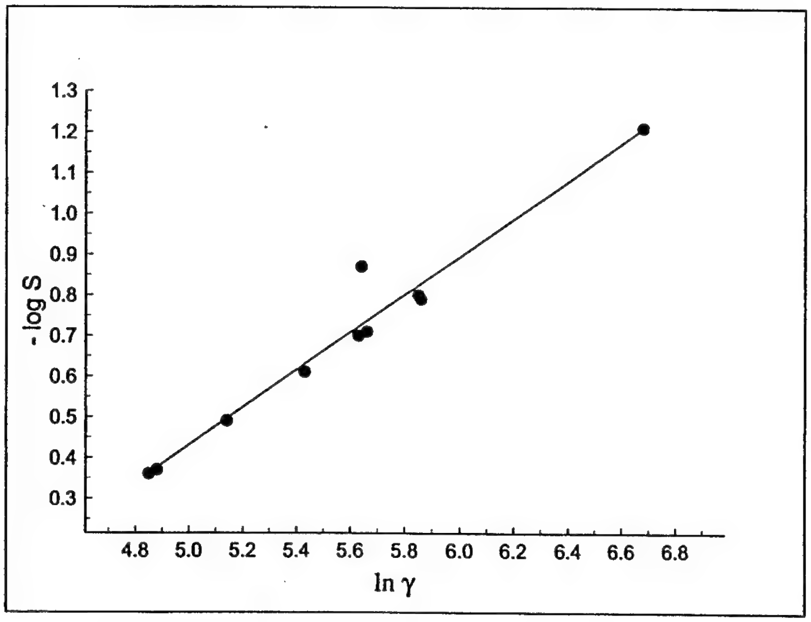


FIGURE 2 The regression between the solubilities $-\log$ S and $\ln\gamma$ for subset of isomers of 1-hexanol.

TABLE IV Comparison of correlation parameters for property-property correlations for the subset of heptanols only. The asterisk (*) indicates the regression in which outlier was removed

Property-property	*r*	*s*	*F*
$\ln \gamma/(-\log S)$	0.9790	0.1163	184.9
* $\ln \gamma/(-\log S)$	0.9987	0.0313	2658.2
b. p./CSA	0.8932	5.6100	31.6
log P/b.p.	0.9484	0.0827	71.5
$\ln \gamma$/b.p.	0.9003	0.2486	34.2

an order of magnitude) of the Fisher ratio for regression including and excluding 2-hexanol.

WEIGHTED PATHS AS DESCRIPTORS

Even though correlations between different properties may vary considerably a single set of well selected molecular descriptors, may nevertheless provide a basis for their regression analysis. This has been already illustrated using a set of the connectivity indices in correlating different physicochemical properties of alkanes [54, 55]. However all previous such studies were based on "fixed" molecular descriptors (topological indices). It is of interest to see how variable molecular topological indices using an optimization procedure to determine the best set of descriptors would describe different molecular properties for the same very sets of compounds.

In Table V we listed the count of smaller paths in alcohols by discriminating C–O bond to which we give weight x. For p_1 this simply increases the count of CC bonds by x, but even this increment may be different for different properties.

RESULTS

We should not be surprised that the weights of paths x vary when we consider different properties even for the same set of compounds. We have seen already that different molecular properties, particularly when focusing attention to isomeric variations, do not correlate at all one with another.

We have previously reported a quite successful correlation for alcohol boiling points when using variable path numbers. In the case of alcohols it was found that optimal weight for CO bond x = 2.2 reduced the standard error to s = 4.82 when path numbers p_1 and p_2 were used as descriptors, and to s = 4.78 when

TABLE V Count of smaller paths in alcohols with CO having weight x

Alcohol	p_1	p_2	p_3	p_4	p_5
1-Butanol	3 + x	2 + x	1 + x	0	0
2M-1-propanol	3 + x	3 + x	2x	0	0
1-Pentanol	4 + x	3 + x	2 + x	1 + x	0
3M-1-butanol	4 + x	4 + x	2 + x	1 + x	0
2M-1-butanol	4 + x	4 + x	2 + 2x	x	0
2-Pentanol	4 + x	3 + 2x	2 + x	1 + x	0
1-Hexanol	5 + x	4 + x	3 + x	2 + x	1 + x
2-Hexanol	5 + x	4 + 2x	3 + x	2 + x	1 + x
3-Hexanol	5 + x	4 + 2x	3 + 2x	2 + x	1
3M-3-pentanol	5 + x	5 + 3x	4 + 2x	1	0
2M-2-pentanol	5 + x	5 + 3x	3 + x	2 + x	0
2M-3-pentanol	5 + x	5 + 2x	3 + 3x	2	0
3M-2-pentanol	5 + x	5 + 2x	4 + 2x	1 + x	0
2,3MM-2-butanol	5 + x	6 + 3x	4 + 2x	0	0
3,3MM-2-butanol	5 + x	7 + 2x	3 + 3x	0	0
4M-2-pentanol	5 + x	5 + 2x	3 + x	2 + 2x	0
1-Heptanol	6 + x	5 + x	4 + x	3 + x	2 + x
2M-2-hexanol	6 + x	6 + 3x	4 + x	3 + x	2 + x
3M-3-hexanol	6 + x	6 + 3x	5 + 2x	3 + x	1
3E-3-pentanol	6 + x	6 + 3x	6 + 3x	3	0
2,3MM-2-pentanol	6 + x	7 + 3x	6 + 2x	2 + x	0
2,3MM-3-pentanol	6 + x	7 + 3x	6 + 3x	2	0
2,4MM-3-pentanol	6 + x	7 + 2x	4 + 4x	4	0
2,2-MM-3-pentanol	6 + x	8 + 2x	4 + 4x	3	0
1-Octanol	7 + x	6 + x	5 + x	4 + x	3 + x
2,2,3MMM-3-pentanol	7 + x	9 + 3x	8 + 4x	3	0
1-Nonanol	8 + x	7 + x	6 + x	5 + x	4 + x

path numbers p_1, p_2 and p_3 were used as descriptors. The above results can be compared with the standard error of 9 °C, obtained by Nikolic, Trinajstić, and Mihalić [56], who considered the Wiener number, the Shultz index, and the valence connectivity index as descriptors. Admittedly these authors considered regressions based on a single descriptor in order to evaluate the relative merits of individual descriptors. Hence, the standard error of 9 °C is not directly comparable to the standard error when one uses two or more descriptors (which can drop to bellow 5 °C). However, if one is interested in obtaining the best regression having statistical significance and giving as small as possible standard error than clearly the procedure based on optimally weighted paths has, as demonstrated, its advantages.

A number of interesting questions can be posed: (1) Does the optimal weight depends on compounds (alcohols) selected? In particular, does it depend on the size of molecules? (2) Does the optimal value of x depends on the number of parameters used? (3) Does the optimal values for x depends on the property considered? Here we will focus on the last two questions. In Table VI we

TABLE VI Dependence of the statistical parameters on the CO bond weight. The optimal value of the weight x is emphasized

(a) Surface cavity area (CSA)

x	r	s	F
−1	0.9645	15.070	205
0	0.9952	5.583	1588
0.3	0.9977	3.865	3330
0.5	**0.9980**	**3.599**	**3842**
0.7	0.9976	3.918	3241
1	0.9964	4.848	2111
1.5	0.9937	6.382	1212
2	0.9913	7.722	868
2.5	0.9893	8.337	704
3	0.9877	8.934	611
3.5	0.9863	9.497	549
4	0.9854	9.734	512
5	0.9838	10.239	461
6	0.9827	10.585	431
7	0.9818	10.835	410
(b) Water solubilities (− log S)			
1	0.9883	0.1653	644
1.5	0.9925	0.1325	1011
2	0.9946	0.1127	1402
2.4	0.9954	0.1038	1655
2.5	0.9955	0.1023	1706
2.6	0.9956	0.1018	1721
3	0.9959	0.0975	1879
3.5	0.9961	0.0961	1932
4	**0.9961**	**0.0960**	**1941**
5	0.9959	0.0981	1855
6	0.9907	0.1011	1749
7	0.9954	0.1039	1655
(c) Octanol -Water partition (log P)			
1	0.9845	0.1369	358
1.5	0.9873	0.1240	439
2	0.9885	0.1183	483
2.25	0.9887	0.1170	498
2.5	0.9889	0.1160	503.1
3	**0.9890**	**0.1156**	**506.3**
3.25	0.9890	0.1157	505.8
3.5	0.9890	0.1158	504.6
4	0.9886	0.1175	491

show the dependence of the statistical parameters r, s, and F on the weight x for each property separately. As we see from Table VI even though we have essentially the same set of compounds the optimal weights vary from property to property displaying dramatic changes. For each property we gave

TABLE VI (*Continued*)

(d) Infinite dilution activity coefficient ($\ln \gamma$)

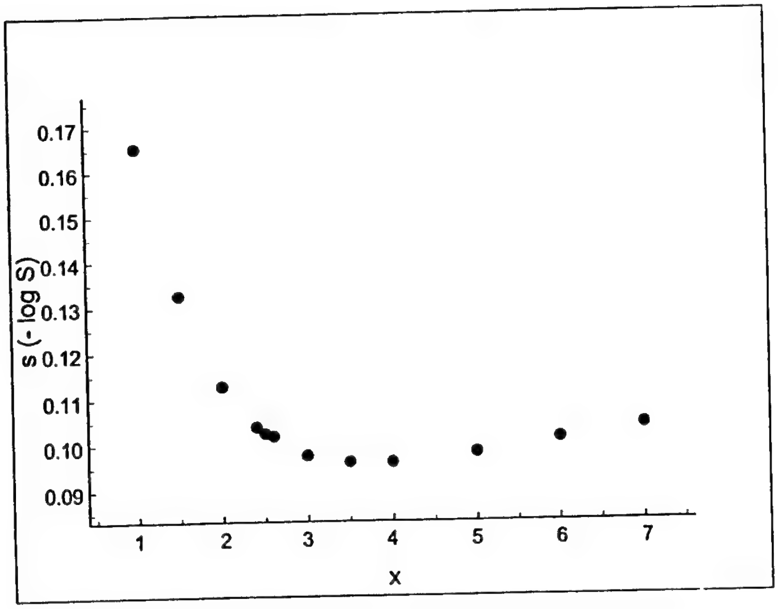
x	r	s	F
1	0.9974	0.4021	2493
2	0.9989	0.2674	5656
3	0.9992	0.2174	8564
4	0.9994	0.1995	10173
5	0.9994	0.1892	11307
6	0.9994	0.1854	11782
7	0.9995	0.1836	12007
8	0.9995	0.1829	12100
9	**0.9995**	**0.1827**	**12124**
10	0.9995	0.1828	12112
12	0.9995	0.1834	12039
15	0.9995	0.1844	11903

the correlation coefficient r, the standard error s, and the Fisher ratio F, as they vary with x, which has been confined to the appropriate domains. In view of relatively small number of molecules in each set (between 38 and 50) we limited the number of descriptors at most three and have used p_1, p_2 and p_3.

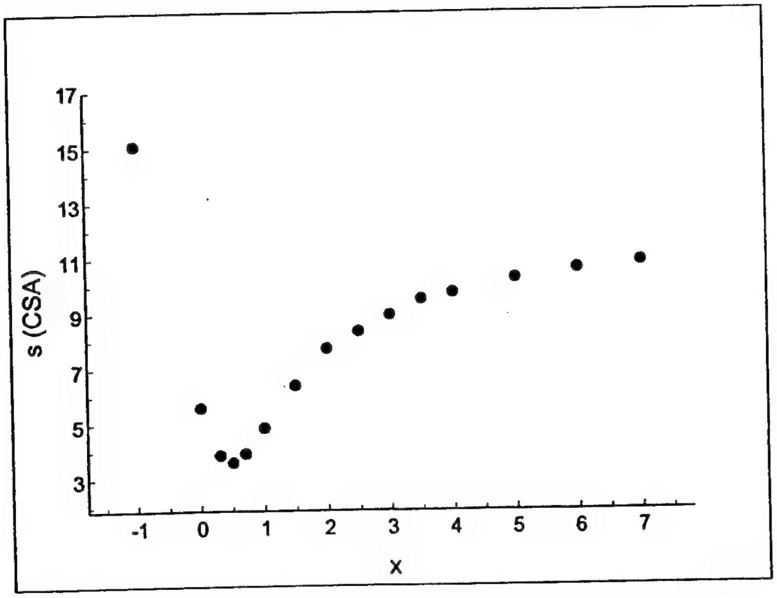
For CSA the best value found for the weight (which is emphasized in Table VI) is: x = 0.5, the value x = 3 is optimal for log P regression, the value x = 4 is optimal for $-\log$ S, and finally the value x = 9 is the optimal value for $\ln \gamma$. These values of x may be compared to x = 2.2 found as the best value for the boiling points of alcohols. Hence, clearly the weight x critically depends on the property considered.

The increase of the weight x means that the role of C–O bond relative to C–C bonds is gaining in the importance. In Figure 3a we have illustrated for the regression of $-\log$ S against the weighted paths p_1, p_2, p_3 the variation of the standard error s against the weight x while in Figure 3b the similar dependence of the standard error s against the weight x is shown for CSA. Both figures show the position of the minimum which corresponds to the optimal weight for x and show a characteristic asymmetric shape of the dependence of s(x) similar in shape to potential curves for a diatomic molecules, or parts of such curves.

Table VII lists the optimal paths p_2 and p_3 for the common 27 alcohols (for which data on all four properties were available) when optimal values of x are selected for each property. The optimal path p_1 are not listed and can be easily derived using expression $p_1 = nCC + x$, where nCC is the number of CC bonds in a molecule. The occurrence of different weights for different properties introduces changes in the relative role of shorter and longer paths for



(a)



(b)

FIGURE 3 (a) Variation of the standard error s against the weight x for the regression of $-\log$ S using weighted paths p_1, p_2, p_3 as descriptors; (b) Variation of the standard error s against the weight x for the regression of CSA using weighted paths p_1, p_2, p_3 as descriptors.

TABLE VII The optimal weighted paths p_2 and p_3 for the five properties of alcohols

Alcohol	$-log\ S$ $x = 4$		*CSA* $x = 0.5$		$log\ P$ $x = 3$		*In* $x = 9$	
1-Butanol	6	5	2.5	1.5	5	4	11	10
2M-1-propanol	7	8	3.5	1	6	6	12	18
1-Pentanol	7	6	3.5	2.5	6	5	12	11
3M-1-butanol	8	6	4.5	2.5	7	5	13	11
2M-1-butanol	8	10	4.5	3	7	8	13	20
2-Pentanol	11	6	4	2.5	9	5	21	11
1-Hexanol	8	7	4.5	3.5	7	6	13	12
2-Hexanol	12	7	5	3.5	10	6	22	12
3-Hexanol	12	11	5	4	10	9	22	21
3M-3-pentanol	17	12	6.5	5	14	10	32	22
2M-2-pentanol	17	7	6.5	3.5	14	6	32	12
2M-3-pentanol	13	15	6	4.5	11	12	23	30
3M-2-pentanol	13	12	6	5	11	10	23	22
2,3MM-2-butanol	18	12	7.5	5	15	10	33	22
3,3MM-2-butanol	15	15	8	4.5	13	12	25	30
4M-2-pentanol	13	7	6	3.5	11	6	23	12
1-Heptanol	9	8	5.5	4.5	8	7	14	13
2M-2-hexanol	18	8	7.5	4.5	15	7	33	13
3M-3-hexanol	18	13	7.5	6	15	11	33	23
3E-3-pentanol	18	18	7.5	7.5	15	15	33	33
2,3MM-2-pentanol	19	14	8.5	7	16	11	34	24
2,3MM-3-pentanol	19	18	8.5	7.5	16	15	34	33
2,4MM-3-pentanol	15	20	8	6	13	16	25	40
2,2-MM-3-pentanol	16	20	9	6	14	16	26	40
1-Octanol	10	9	6.5	5.5	9	8	15	14
2,2,3MMM-3-pentanol	22	24	11.5	10	19	20	36	44
1-Nonanol	11	10	7.5	6.5	10	9	16	15

different structures. Consider for example 2-methyl-1-butanol and 2-pentanol (of Table II). When $x = 0.5$ (the optimal value for CSA) the quotient p_2/p_3 for 2-methyl-1-butanol and 2-pentanol are not very different, 4.5/3 and 4/2.5 respectively. In contrast when $x = 4$ (optimal value for $-\log$ S) the quotient p_2/p_3 for 2-methyl-1-butanol and 2-pentanol are very different, 8/10 and 11/6 respectively. The standard topological indices lack the flexibility to adjust similarly to such demand dictated by diversity of properties.

In Table VIII we listed the calculated properties and the residuals of the regression as obtained for the common $n = 27$ alcohols. For all the four properties all the caclulated values are within two standard deviations, except in the case of SCA where highly branched 2, 2, 3-trimethyl-3-pentanol shows large residual. The calulaterd CSA is found too small: 324.27, the reported experimental value is however 335.2. By discarding this point as an outlier the standard error dropps to 3.124. The regression equations are listed in Table IX.

TABLE VIII Calculated properties of alcohols. We displayed only the results for alcohols of Table II, but the calculations were based on all alcohols for which data were available

Alcohol	*−log S**	*Residual*	*CSA*	*Residual*
1-Butanol	0.035	−0.035	270.19	1.91
2M-1-propanol	−0.083	0.073	262.94	0.86
1-Pentanol	0.620	−0.040	301.73	2.17
3M-1-butanol	0.538	−0.028	291.93	−0.53
2M-1-butanol	0.490	−0.030	289.36	0.04
2-Pentanol	0.293	−0.013	296.83	−0.93
1-Hexanol	1.205	0.005	333.28	2.42
2-Hexanol	0.878	−0.008	328.38	−0.68
3-Hexanol	0.830	−0.030	325.81	−0.51
3M-3-pentanol	0.410	−0.050	305.98	−0.18
2M-2-pentanol	0.470	0.020	313.67	0.63
2M-3-pentanol	0.700	0	313.44	0.86
3M-2-pentanol	0.736	−0.026	310.88	0.42
2,3MM-2-butanol	0.329	0.042	296.18	5.03
3,3MM-2-butanol	0.537	0.073	303.86	3.64
4M-2-pentanol	0.797	−0.007	318.57	−3.67
1-Heptanol	1.790	0.020	364.83	2.67
2M-2-hexanol	1.055	0.015	345.21	0.89
3M-3-hexanol	0.995	−0.015	337.52	0.18
3E-3-pentanol	0.934	−0.104	329.83	−5.43
2,3MM-2-pentanol	0.901	−0.031	322.59	1.21
2,3MM-3-pentanol	0.853	−0.013	320.02	1.76
2,4MM-3-pentanol	1.155	0.065	332.62	−0.92
2,2-MM-3-pentanol	1.074	0.076	322.81	3.29
1-Octanol	2.375	−0.025	396.37	3.03
2,2,3MMM-3-pentanol	1.214	0.056	324.27	10.93
1-Nonanol	2.960	0.040	427.92	3.28

COMPARISON WITH MRA FROM OTHER SOURCES

Comparison between different regression results are primarily of interest because they can point to dominant and the most relevant molecular descriptors for properties studied. When such descriptors are identified they can assist in revising or refining molecular models for compounds considered. The standard error is likely to point to most useful regression if the accuracy of the prediction is the only criteria considered. However, the standard error important as it is, is not necessarily the only parameter of interest in structure-property-activity studies. Equally important, or even more important, may be the structural meaning of the descriptors used as they can facilitate not only an improvement of the model used but also may offer a better insight into our understanding of the structure-property relationship, even though structure-property correlation does not invoke causal relationship.

A strict comparison between different regression results is only possible if the two studies use the same experimental data on the same set of compounds with

TABLE VIII (*Continued*)

Alcohol	*log P*	*Residual*	*ln* γ	*Residual*
1-Butanol	0.802	0.078	4.018	−0.098
2M-1-propanol	0.703	−0.093	3.829	0.061
1-Pentanol	1.310	0.091	5.372	−0.082
3M-1-butanol	1.243	−0.103	5.281	0.059
2M-1-butanol	1.194	−0.054	5.171	−0.091
2-Pentanol	1.109	0.031	4.556	0.014
1-Hexanol	1.817	0.213	6.726	−0.046
2-Hexanol	1.617	−0.007	5.910	−0.270
3-Hexanol	1.568	0.042	5.799	0.051
3M-3-pentanol	1.285	0.106	4.880	−0.030
2M-2-pentanol	1.349	0.041	5.003	0.137
2M-3-pentanol	1.452	−0.042	5.598	0.032
3M-2-pentanol	1.485	−0.075	5.696	−0.036
2,3MM-2-butanol	1.218	−0.048	4.789	0.091
3,3MM-2-butanol	1.319	−0.129	5.416	0.014
4M-2-pentanol	1.550	−0.140	5.819	0.041
1-Heptanol	2.324	0.016	8.080	0.010
2M-2-hexanol	1.857	0.013	6.357	0.133
3M-3-hexanol	1.792	0.078	6.234	0.056
3E-3-pentanol	1.727	0.143	6.111	−0.171
2,3MM-2-pentanol	1.725	−0.055	6.131	−0.111
2,3MM-3-pentanol	1.660	0.010	6.020	−0.060
2,4MM-3-pentanol	1.844	−0.134	6.750	0.070
2,2-MM-3-pentanol	1.778	−0.088	6.660	0.000
1-Octanol	2.832	0.008	9.434	0.126
2,2,3MMM-3-pentanol	1.969	0.021	7.161	−0.211
1-Nonanol	3.339	−0.189	10.788	0.212

TABLE IX The regression equations

Property	p_1	p_2	p_3	*Constant*
CSA	46.4797	−9.8066	−5.1272	139.7160
− log S	0.6660	−0.0802	−0.0115	−4.0677
log P	0.5904	−0.0668	−0.0162	−2.3415
ln γ	1.4571	−0.0907	−0.0123	−10.8897

the same number of descriptors. This is rarely the case, because between two studies novel data may be available and is likely to be included in more recent work. In addition different authors may have their own preferences for selecting and testing descriptors using larger set of compounds that allow increased number of descriptors. Our comparison here is of such a kind because Cao and Li [52] who reported MRA on water solubility, surface cavity area, and log P included in their set of alcohols also alkanes and cyclo-alkanes. Similarly Mitchell and Jurs [53] besides alcohols included a variety of organic compounds having other heteroatoms (halogens, nitrogen). As we will see for our results the standard error has been dramatically decreased, in comparison with the above mentioned results, except for correlation for log P where the improvement is significant, but not dramatic. In view of the differences in the size of samples and the diversity of compounds it should not be surprising that we get smaller standard error than others. What is surprising is by how much we have reduced the standard error when using optimized descriptors.

Here are listed r and s for the property studies as reported in ref. [52, 53] and in this work:

Property	*n*	*N*	*r*	*s*	*Ref.*
CSA	69	2	0.9954	5.20	52
$-\log S$	60	2	0.994	0.167	52
$\log P$	54	3	0.992	0.124	52
$\ln\gamma$	296	12	0.978	0.753	53
$\ln\gamma$	271	12		0.376	53
$\ln\gamma$	193	12	0.967	0.559	53

Property	*n*	*N*	*r*	*s*	*F*	*Ref.*
CSA	50	3	0.9980	3.599	3842	this work
SCA*	49	3	0.9985	3.124	5104	this work
$-\log S$	50	3	0.9961	0.0960	1941	this work
$-\log S^*$	48	3	0.9978	0.0713	3324	this work
$\log P$	50	3	0.9890	0.1156	506	this work
$\ln\gamma$	50	3	0.9995	0.1827	12124	this work

Here n is the size of the sample (structures) and N is the number of parameters (descriptors) used in the regressions, while F is Fisher ratio.

CONCLUDING REMARKS

We have outlined a novel way of deriving powerful structure-property models. We consider assigning to shorter paths in molecules variable weight x, to be determined during the regression analysis so that one obtains the smallest standard error for correlation considered. Even though the approach has been demonstrated on several physico-chemical properties of simple chemical structures, it is general and applies to analysis of properties of more complex chemical compounds. The advantage of the outlined approach is that it yields regressions accompanied with considerably smaller standard error than are given by similar studies using standard molecular descriptors. The "flexibility" of the molecular descriptor, such as weighted paths used in this study, in contrast to the "fixed" molecular descriptors, which are numerically determined one molecular structure is known, makes it possible to describe different properties of a same set of compounds by the same kind of descriptors. As can be seen from the illustration given by analysis of several properties of smaller alcohols different properties may require different weighting factors. This suggests that methods in which prescribed modification of topological indices are assumed in order to describe heteroatoms, such as for example the valence connectivity indices of Kier and Hall, have inherent limitations, in that they may be suitable for some molecular properties but less suitable for others. Indeed, several authors have reported correlations for compounds involving heteroatoms for which a simple connectivity index gives a better regression that the corresponding valence connectivity index.

Acknowledgment

This work was supported in part by Ministry of Science and Technology of Slovenia through grant J1-8901.

References

[1] Hotelling, H. (1933). Analysis of complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–489.

[2] Wold, S. and Sjostrom, M. (1977). In Chemometrics: Theory and Applications (ACS Symp. Ser. No. 52), Kowalski, B. R. (Ed.), Am. Chem. Soc., Washington, D. C., p. 243.

[3] Wold, S., Sjöström, M. and Eriksson, L. (1998). Partial Least Squares projections to latent Structures (PLS) in chemistry. In: *Encyclopedia of Computational Chemistry*, (von Schleyer, R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F. and Schreiner, P. R. Eds.). John Wiley & Sons, Chichester, England, pp. 2006–2021.

[4] Zupan, J. (1998). In: Neural networks in chemistry. *Encyclopedia of Computational Chemistry*. (von Schleyer, R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F. and Schreiner, P. R. Eds.). John Wiley & Sons, Chichester, England, **3**, 1813–1827.

[5] Malinowski, E. R. (1991). *Factor analysis in Chemistry*. John Wiley & Sons, New York.

[6] Randić, M. and Seybold, P. G. (1993). Molecular shape as a critical factor in structure-property-activity studies. *SAR and QSAR in Environ. Res.*, **1**, 77–85.

[7] Randić, M. (1992). On the representation of molecular graphs by basis graphs. *J. Chem. Inf. Comput. Sci.*, **32**, 57–69.

[8] Kier, L. B. and Hall, L. H. (1976). *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York.

[9] Randić, M. (1991). Orthogonal molecular descriptors. *New J. Chem.*, **15**, 517–525.

[10] Randić, M. (1991). Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.*, **31**, 311–320.

[11] Randić, M. (1993). Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.*, **14**, 363–370.

[12] Randić, M. (1994). Curve-fitting paradox. *Int. J. Quant. Chem: Quant. Biol. Symp.*, **21**, 215–225.

[13] Randić, M. (1996). Orthosimilarity. *J. Chem. Inf. Comput. Sci.*, **36**, 1092–1097.

[14] Pogliani, L. (1996). Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J. Phys. Chem.*, **100**, 18065–18077.

[15] Šoškić, M., Plavšić, D. and Trinajstić, N. (1996). Link between orthogonal and standard multiple regression models. *J. Chem. Inf. Comput. Sci.*, **36**, 829–832.

[16] Amić, D., Davidović-Amic, D. and Trinajstić, N. (1995). Calculation of retention times of anthocyanins with orthogonalized topological indices. *J. Chem. Inf. Comput. Sci.*, **35**, 136–139.

[17] Hall, L. H. MOLCONN software.

[18] Basak, S. C., Harris, D. K. and Magnuson, V. R. POLLY (version 2.3), Copyright of the University of Minnesota.

[19] Katritzky, A. R., Lobanov, V. S. and Karelson, M. (1995). QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.*, **24**, 279–287.

[20] Randić, M. (1997). On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.*, **37**, 672–687.

[21] Randić, M., Novič, M. and Vračko, M. *Molecular Descriptors, New and Old*, Lecture Notes in Chemistry (submitted)

[22] Kier, L. B. and Hall, L. H. (1976). Molecular connectivity VII. Specific treatment of heteroatoms. *J. Pharm. Sci.*, **65**, 1806–1809.

[23] Kupchik, E. J. (1989). General treatment of heteroatoms with the Randic molecular connectivity index. *Quant. Struct. —Act. Relat.*, **8**, 98–103.

[24] Randić, M. (1991). On computation of optimal parameters for multivariate analysis of structure-property relationship. *J. Comput. Chem.*, **12**, 970–980.

[25] Spialter, L. (1963). The atom connectivity matrix (ACM) and its characteristic polynomial (ACMP): A new computer oriented chemical nomenclature. *J. Am. Chem. Soc.*, **85**, 2012–2013.

[26] Randić, M. (1991). Novel graph theoretical approach to heteroatom in quantitative structure-activity relationship. *Chemometrics & Intel. Lab. Syst.*, **12**, 970–980.

[27] Hansch, C. and Leo, A. (1995). *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*. ACS Professional Reference Book, ACS Washington, DC, p. 557.

[28] Randić, M. and Dobrowolski, J. Cz. (1998). Optimal molecular connectivity descriptors for nitrogen containing molecules. *Int. J. Quant. Chem: Quant. Biol. Symp.*, **70**, 1209–1215.

[29] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.

[30] Hosoya, H. (1971). Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Japan*, **44**, 2332–2339.

[31] Randić, M., Acta Chim. Slovenica (submitted)

[32] Randić, M. (1997). Linear combinations of path numbers as molecular descriptors. *New J. Chem.*, **21**, 945–951.

[33] Randić, M. and Pompe, M. (1999). On characterization of CC double bond in alkenes. *SAR QSAR Environ. Res.*, **10**,

[34] Platt, J. R. (1947). Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.*, **15**, 419–420.

[35] Randić, M. (1978). Fragment search in acyclic structures. *J. Chem. Inf. Comput. Sci.*, **18**, 101–107.

[36] Randić, M. and Wilkins, C. L. (1979). On a graph theoretical basis for ordering of structures. *Chem. Phys. Lett.*, **63**, 332–336.

[37] Randić, M. and Wilkins, C. L. (1979). Graph-based fragment search in polycyclic structures. *J. Chem. Inf. Comput. Sci.*, **19**, 23–31.

[38] Randić, M., Brissey, G. M., Spencer, R. G. and Wilkins, C. L. (1979). Search for all self-avoiding paths for molecular graphs. *Comput. Chem.*, **3**, 5–13.

[39] Randić, M. and Wilkins, C. L. (1979). Graph theoretical study of structural similarity in benzomorphans. *Int. J. Quant. Chem: Quant. Biol. Symp.*, **6**, 55–71.

[40] Randić, M. and Wilkins, C. L. (1979). Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Chem. Phys.*, **83**, 1525–1540 (additions and corrections: *J. Chem. Phys.*, (1980), **84**, 2090.

[41] Randić, M. (1979). Characterization of atoms, molecules, and classes of molecules based on path enumerations. *MATCH*, **7**, 3–60.

[42] Randić, M. and Wilkins, C. L. (1980). Graph theoretical analysis of molecular properties. Isomeric variations in nonanes. *Int. J. Quant. Chem.*, **18**, 1005–1027.

[43] Wilkins, C. L. and Randić, M. (1980). A graph theoretical approach to structure-property and structure-activity correlations. *Theor. Chim. Acta*, **58**, 45–68.

[44] Randić, M. (1984). On molecular identification numbers. *J. Chem. Inf. Comput. Sci.*, **24**, 164–175.

[45] Randić, M. (1984). Nonempirical approach to structure-activity studies. *Int. J. Quant. Chem: Quant. Biol. Symp.*, **11**, 137–153.

[46] Randić, M. (1985). Graph theoretical approach to structure-activity studies. Search for optimal antitumor compounds In: *Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogens* (Rein, R. Ed.). Alan R. Liss, Publ., pp. 309–318.

[47] Randić, M., Jerman-Blažič, B., Grosman, S. C. and Rouvray, D. H. (1986). A rational approach to the optimal drug design. *Math. Modeling*, **8**, 571–582.

[48] Randić, M., Jerman-Blažič, B., Rouvray, D. H., Seybold, P. G. and Grosman, S. C. (1987). The search for active substructure in structure-activity studies. *Int. J. Quant. Chem: Quant. Biol. Symp.*, **14**, 245–260.

[49] Randić, M., Grosman, S. C., Jerman-Balzic, B., Rouvray, D. H. and El-Basil, S. (1988). An approach to modeling the mutagenicity of nitroarenes. *Math. Comput. Modeling*, **11**, 837–842.

[50] Kier, L. B., Murray, W. J., Randić, M. and Hall, L. H. (1976). Molecular connectivity V. Connectivity series concept applied to density. *J. Pharm. Sci.*, **65**, 1226–1230.

[51] Randić, M. and Basak, S. C. (1999). Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.*, **39**, 261–266.

[52] Cao, C. and Li, Z. (1998). Molecular polarizability. 1. Relationship to water solubility of alkanes and alcohols. *J. Chem. Inf. Comput. Sci.*, **38**, 1–7.

[53] Mitchell, B. E. and Jurs, P. C. (1998). Prediction of infinite dilution activity coefficients of organic compounds in aqueous solution from molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 200–209.

[54] Needham, D. E., Wei, I.-C. and Seybold, P. G. (1988). Molecular modeling of the physical properties of the alkanes. *J. Am. Chem. Soc.*, **110**, 4186–4194.

[55] Randić, M. and Trinajstić, N. (1993). Viewpoint 4 - Comparative structure-property studies: The connectivity basis. *J. Mol. Struct. (Theochem)*, **284**, 209–221.

[56] Nikolić, S., Trinajstić, N. and Mihalić, Z. (1993). Molecular topological index. *J. Math. Chem.*, **12**, 251–264.

APPENDIX 1.9 On 3-D graphical representation of DNA primary sequences and their numerical characterization

On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization

M. Randić,[†,‡,§,#] M. Vračko,[†] A. Nandy,[||] and S. C. Basak*[,§]

National Institute of Chemistry, 1001 Ljubljana, POB 3430, Slovenia, Ames Laboratory - DOE,
Iowa State University, Ames, Iowa 50011, Natural Resources Research Institute,
University of Minnesota at Duluth, Miller Trunk Highway, Duluth Minnesota 55811, and
Computer Division, Indian Institute of Chemical Biology, Calcutta, India

Received April 9, 2000

In this article we (1) outline the construction of a 3-D "graphical" representation of DNA primary sequences, illustrated on a portion of the human β globin gene; (2) describe a particular scheme that transforms the above 3-D spatial representation of DNA into a numerical matrix representation; (3) illustrate construction of matrix invariants for DNA sequences; and (4) suggest a data reduction based on statistical analysis of matrix invariants generated for DNA. Each of the four contributions represents a novel development that we hope will facilitate comparative studies of DNA and open new directions for representation and characterization of DNA primary sequences.

INTRODUCTION

With rapid reporting of DNA sequences derived with automated DNA sequencing techniques the problem of processing such information became acute. Usual representation of the primary sequence DNA is that of a string of letters A, G, C, T, which signify the four nucleic acid bases adenine, guanine, cytosine, and thymine, respectively. Such sequences can be very long, and even the segments of interests when comparing DNA of different species can be quite lengthy. In Table 1 we listed DNA of human β globin gene. Its length is 1424, and its first exon already involves 92 bases. Comparison of such primary sequences, and even their fragments having less than 100 bases, could be quite difficult for several reasons. Consider the list of the first exon of the β globin gene for eight different species shown in Table 2. They all look similar, but at the same time they are all sufficiently different. How similar or how different they are may depend on how such strings of letters are encoded or characterized. The standard procedures consider differences between strings due to deletion–insertion, compression–expansion, and substitution of the string elements.[1–9] These approaches have been applied to a variety of problems, from the error correcting codes in which Levenshtein has introduced metrics for string comparisons[1] to comparison of DNA sequences, comparison of protein sequences, and applications in quantitative structure–activity relationship (QSAR).[8,9] Such approaches, that have been hitherto widely used, are computer intensive.

We have recently proposed an alternative approach for comparison of sequences that is based on characterization of DNA by ordered sets of *invariants* derived for DNA sequence, rather than by a direct comparison of DNA sequences themselves. This is analogous to the use of graph invariants (topological indices) for characterization of molecules rather than use of information on their geometry and types of atoms involved. An important advantage of a characterization of structures (be it molecule or DNA) by invariants, as opposed to use of codes, is the simplicity of the comparison of numerical sequences based on invariants. The price paid is a loss of information on some aspects of the structure that accompany any characterization based on invariants. The loss of information, however, can be in part reduced by use of a larger number of descriptors (invariants), as has been well illustrated in SAR and QSAR based on mathematical descriptors for molecules.[10–12]

Graphical representations of DNA that have been developed within the past few years[13–15] offer a route to one such condensation of information coded by DNA primary sequence into a set of invariants. In Figure 1 we show few graphical representations of selected DNA as reported by Nandy.[16] The graphs are obtained by assigning to the four directions associated with the positive and the negative *x, y* axes the four nucleic acid bases A, G, C, T, such that A and T correspond to the negative *x, y* axes, respectively, and G and C correspond to the positive *x* and *y* axes, respectively. An advantage of graphical representations of DNA is that it allows visual comparisons which are easier to make. One should, however, be aware of a loss of information inherent in such graphical representations. One of the limitations is that graphical form shows the "path" of the "travel" along the primary sequence but not the "history" of the travel. *Hence, we do not know when what parts of the graphical path were retraced.* At the top of Figure 2 we show a graphical representation of the first exon of the human β globin gene, at a higher magnification. The rest of Figure 2 shows the first exon of β globin gene of several other species for comparison. As we can see upon inspection qualitative

[*] Corresponding author phone: (218)720-4328; fax: (218)720-4330; e-mail: sbasak@nrri.umn.edu.

[†] National Institute of Chemistry.

[‡] Iowa State University.

[§] University of Minnesota at Duluth.

[||] Indian Institute of Chemical Biology.

[#] On leave from Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311.

10.1021/ci000034q CCC: $19.00
Published on Web 08/15/2000

Table 1. DNA of Length 1424 Listing Nucleic Bases in Human Beta Globin Gene[a]

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGG
TATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCA
TGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTC
TGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCTTGG
ACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGT
TATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGC
CTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTT
CAGGGTGAGTCTATGGGACCCTTGATGTTTTCTTTCCCCTTCTTTTCTATG
GTTAAGTTCATGTCATAGGAAGGGGAGAAGTAACAGGGTACAGTTTAG
AATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCG
TTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAAT
TCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATG
CCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATACATTAAG
TAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTG
GAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTC
TATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAATG
ATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGAT
AATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTG
CATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGC
TACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTC
TTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC
ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
TATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAATG
ATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGAT
AATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTG
CATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGC
TACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTC
TTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCC
ATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA

[a] ID HSHBB – beta globin gene sequence extract: exons: 1–92, 223–445, 1296–1424; introns: 93–222, 446–1295. SQ Hshbb.MK1 -- segment from 62205 to 63628 of HSHBB.

Table 2. First Exon of Beta Globin Gene for Eight Species Labeled A–H

A	human beta-globin	92 bases
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG		
GCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG		
B	goat alanine beta-globin	86 bases
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGG		
TGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG		
C	opossum beta-hemoglobin beta M-gene	92 bases
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGT		
CTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG		
D	gallus gallus beta globin	92 bases
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGG		
GCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG		
E	lemur beta-globin	92 bases
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGG		
GCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG		
F	mouse beta-a-globin	93 bases
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGG		
CAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG		
G	rabbit beta-globin	90 bases
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGG		
GCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC		
H	rat beta-globin	92 bases
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGG		
GAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG		

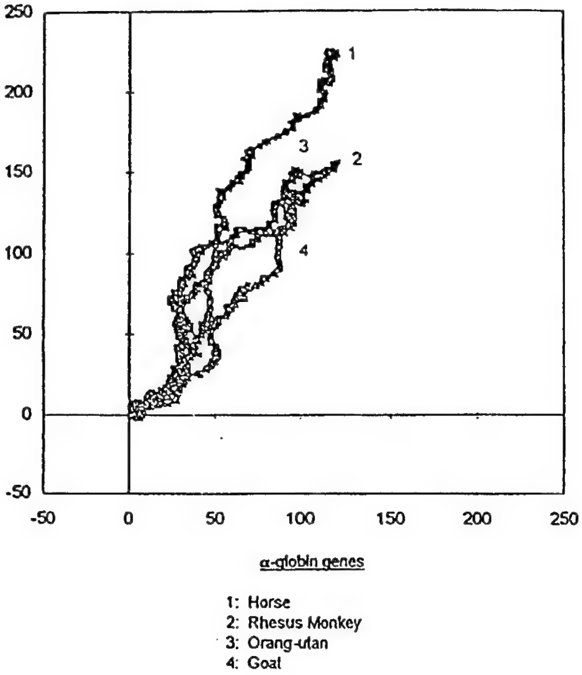


Figure 1. Few graphical representations of selected DNA that Nandy and collaborators developed.

similarities and differences between exons of different species are immediately apparent.

Mathematical curves can be represented in the form f(x, y) = 0, which corresponds to graphical projections of DNA of Figure 2, and in a parametric form $x = x(t)$ and $y = y(t)$. Clearly there is a loss of information in going from a parametric representation of a curve $x = x(t)$ and $y = y(t)$ to the analytical representation of the same curve. The f(x, y) = 0 only represents the path, while the former, if the parameter t is interpreted as time, gives the history of the movement over the path. Equally, there is loss of information when a a spatial curve is represented by its projection in the (x, y) plane (or any other plane). Hence, two routes for an improvement of graphical representations of DNA sequences appear possible: (1) to consider representation analogous to parametric representation of mathematical curves and (2)

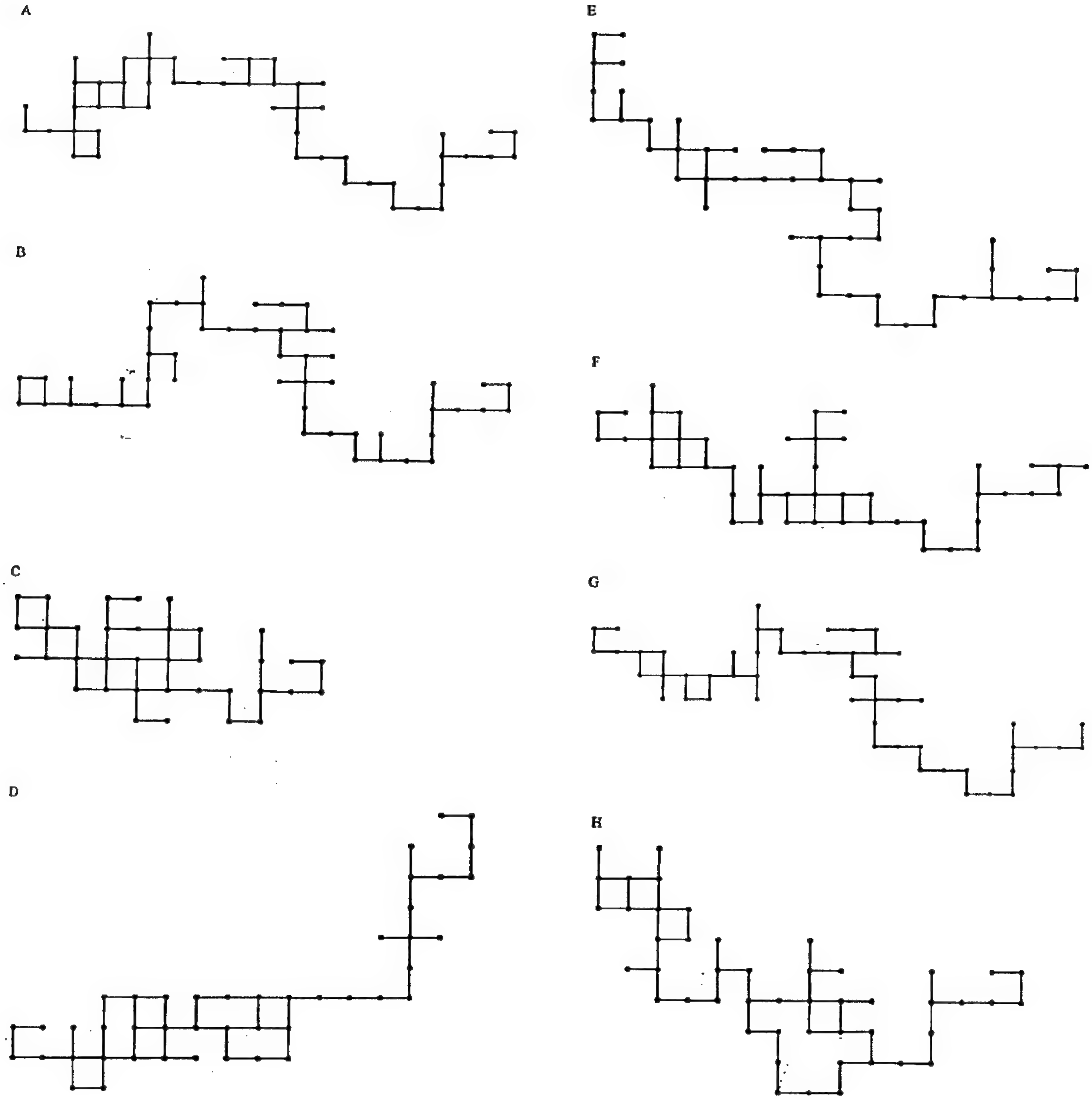


Figure 2. Graphical representations of the first exon of the human beta globin gene (top), a "detail" of Figure 1 and the remaining beta globin genes of Table 2.

to consider graphical representation of DNA sequence with "path" which is in traced in 3D space, rather than a plane. In this paper we will limit our attention to this latter problem. We will then describe a scheme which generates for a graphical spatial representation of DNA a numerical matrix. Once we arrive at a matrix representation of DNA we will search for suitable matrix invariants to be used for characterization of DNA. Finally we will consider possible condensation of derived numerical characterization of DNA in a more compact format.

3-D REPRESENTATION OF DNA PRIMARY SEQUENCE

Two-dimensional representation of DNA developed by Nandy[4] assigned to the four directions defined by the positive and the negative x and y coordinate axes to the four nucleic bases so that A and G are associated with the x-axis and C and T with the y axis. This assignment of directions differs from the assignment considered by Leong and Morgenthaler,[14] who take a move to the right to correspond to A, a move to the left is C, an upward move is a T, and a downward move is G.

The nonequivalent directions are created after assignments of the first base because then there remains only one site that is *opposite* to the already selected direction; the other two sites are at *lateral* positions. If we could have three *equivalent* directions after the first assignment we would avoid considering the multiplicity of alternatives (projections). This is possible by using the directions defined by vertices of a regular tetrahedron. When looking from its center all the four directions toward the four vertices are equivalent, hence after selecting one direction the three directions remain equivalent. Hence, we will assign to four nucleic acid bases the four directions associated with the regular tetrahedron. To specify directions we will place the origin of the Cartesian (x, y, z) coordinate system in the center of a cube (Figure 3) so that the four corners of the cube,

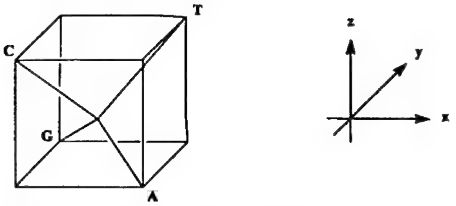


Figure 3. The tetrahedral directions assigned to A, G, C, T nucleic bases.

Table 3. Cartesian 3-D Coordinates for Initial Part of the Sequence of DNA Nucleic Bases of the First Exon

		x	*y*	*z*			*x*	*y*	*z*
1	A	+1	−1	−1	15	T	−1	+1	+1
2	T	+2	0	0	16	C	−2	0	+2
3	G	+1	+1	−1	17	C	−3	−1	+3
4	G	0	+2	−2	18	T	−2	0	+4
5	T	+1	+3	−1	19	G	−3	+1	+3
6	G	0	+4	−2	20	A	−2	0	+2
7	C	−1	+3	−1	21	G	−3	+1	+1
8	A	0	+2	−2	22	G	−4	+2	0
9	C	−1	+1	−1	23	A	−3	+1	−1
10	C	−2	0	0	24	G	−4	+2	−2
11	T	−1	+1	+1	25	A	−3	+1	−3
12	G	−2	+2	0	26	A	−1	0	−4
13	A	−1	+1	−1	27	G	−3	+1	−5
14	C	−2	0	0	28	T	−2	+2	−2

which define the tetrahedral directions, have the coordinates (+1, −1, −1), (−1, +1, −1), (−1, −1, +1), and (+1, +1, +1). To each tetrahedral direction we assign one nucleic base as follows:

$$(+1,-1,-1) \rightarrow \mathrm{A}$$

$$(-1,+1,-1) \rightarrow \mathrm{G}$$

$$(-1,-1,+1) \rightarrow \mathrm{C}$$

$$(+1,+1,+1) \rightarrow \mathrm{T}$$

The particular assignment is arbitrary, but this has no significance since all directions are equivalent. To obtain the spatial path associated with the DNA sequence, we move in *x*, *y*, *z* space in the direction that the above assignments dictates. Consider the beginning of the first exon of Table 1:

A T G G T G C A....

The first point of the spatial curve is at point (+1, −1, −1) which belongs to A, so directed from the origin. From that point we move in the direction assigned to T, (+1, +1, +1), which means that all the three coordinates of the position A, (+1, −1, −1), have to be increased by +1. We arrive then at the point (+2, 0, 0) as the location of T. From here we move in the direction defined by (−1, +1, −1) assigned to G telling that the first and the third coordinates have decreased while the second coordinate has increased. This leads to point (+1, +1, −1) as the location of G. Continuing in the direction of G we have again to decrease *x* and *z* (the first and the third coordinates) and to increase *y* (the second coordinate). Thus we come to the point (0, +2, −2). The process continues, each time we algebraically add the (*x*, *y*, *z*) coordinates of the new point to that of the last point. Continuation of this process is illustrated in Table 3 for the two dozen initial nucleic bases of the first exon. In Figure 4 we show a portion of 3-D graphical representation of DNA of Table 1.

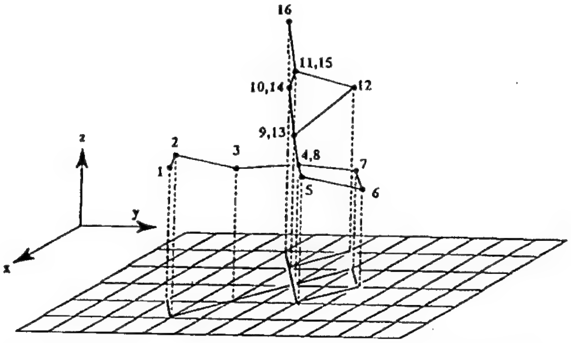


Figure 4. Portion of 3-D graphical representation of DNA of Table 1.

NUMERICAL CHARACTERIZATION OF SPATIAL REPRESENTATION OF DNA

An important advantage of graphical representations of DNA, both 2-D and 3-D, is the possibility to derive numerical characterizations for such mathematical objects. One way to arrive at numerical characterization of DNA is to associate with its graphical representation given by a curve in the space (or a plane) a matrix. Once we have a matrix we can use matrix invariants arrive at various numerical descriptors, rather than the visual description of the DNA sequence. This is analogous to the use of matrices associated with molecular graphs or molecular structure as a source for construction of topological indices rather than using molecular models (such as "sticks-and-balls or "space-filling" models) for their representation.[10]

Formally, there is no difference between a graphical "sequence chain" (in 2-D or 3-D space) or an actual polymer ("atom chain") in the space. Hence, we can transfer mathematical methods used for the characterization of molecules in structure−property and the structure−activity studies to numerical characterization of 3-D representations of the primary DNA sequence. This has been considered recently by Randić and collaborators[17] for 2-D graphical representation of DNA.

We should mention that one can also arrive at numerical descriptors that may be specific and sensitive to graphical form of a DNA without necessarily resorting to matrices. Thus, for example, Raychaudhury and Nandy[18] considered several geometrical parameters of DNA curves, such as, for example, end-to-end distance as DNA descriptors. Matrices, however, offer additional descriptors and richer characterization and can be manipulated by a computer, and one can take other advantages of linear algebra, rather than being confined to ordinary geometry.

Search for novel descriptors may be an endless project, just as this has been the case with mathematical descriptors that continue to be constructed for molecules. However, the art is in finding *useful* descriptors, and those that have plausible structural interpretation, at least within the model considered. Matrices have an additional advantage: they allow one to construct additional matrices by combining elements of different matrices as components. In this way one can arrive at additional descriptors for DNA. In this report we will confine our interest particularly to the graph

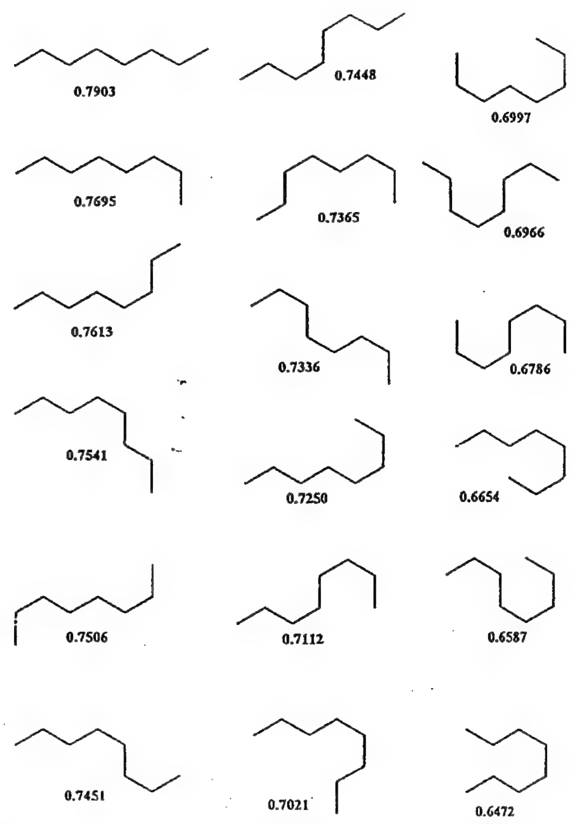


Figure 5. Conformations of eight-atom chain embedded on a graphite lattice ordered according to decreasing values of the leading eigenvalue of **D/D** matrix.

theoretical distance matrix and the Euclidean distance matrix for characterization of graphical forms of DNA.

MATRICES INVOLVING DISTANCES

The input information in a graph distance matrix[19,20] is solely confined to the information on the connectivity of the structure (system). However, when a graph is embedded in a space it assumes a fixed geometry. Then, in addition to the graph theoretical distance between a pair of vertices, we can also compute the Euclidean distances between the same pair of vertices. The Euclidean and the graph theoretical distances can be combined into a single distance/distance matrix by taking the quotient of the corresponding matrix elements.[21,22] Collection of such quotients for all pairs of vertices leads to the so-called **D/D** matrix. Matrices constructed in this way proved very promising as a tool for characterization of structures embedded in 3-D space. The normalized leading eigenvalue λ_1/n of a **D/D** matrix offers a measure of the degree of folding of a chainlike structure or a curve. In Figure 5 we illustrated configurations of an eight-atom C_8 chains embedded on a graphite lattice. Under each skeleton is given the normalized λ_1/n of **D/D** matrix. As we see the largest eigenvalue ($\lambda_1/8 = 0.7903$) is associated with the least bent *all-trans* configuration of C_8, and the smallest eigenvalue ($\lambda_1/8 = 0.6472$) belongs to the highly folded isomer TCCCT. T and C label stand for trans and cis conformations of three consecutive CC bonds (consult Table 4 for structures belonging to different labels). For chains of seven CC bonds even a smaller eigenvalue than 0.6472 is possible. It belongs to the hypothetical *all-cis* configuration CCCCC, the projection of which on hexagonal lattice gives a regular hexagon. In this structure the first and the last CC bond of C_8 would overlap, giving for $\lambda_1 = 4.6388$, which when normalized becomes $\lambda_1/8 = 0.5798$. The relative magnitudes of λ_1/n and the shape of corresponding conformations fully supports the interpretation of the normalized eigenvalue of **D/D** matrices as an index of the folding of a structure.

Table 4. Leading Eigenvalues for **D/D** Matrices of Eight-Atom Chains Embedded on a Graphite Lattice and the Leading Eigenvalues of the Corresponding Line Adjacency Matrices

conformer	λ_1/n of D/D matrix	λ_1/n of line adjacency matrix
TTTTT	0.7903	0.8571
TTTTC	0.7695	0.7191
TTTCT	0.7613	0.5916
TTCTT	0.7541	0.5208
CTTTC	0.7506	0.5858
TTCTC	0.7451	0.4688
TCTCT	0.7448	0.4019
TCTTC	0.7365	0.4748
CTCTC	0.7336	0.3836
TTTCC	0.7250	0.5793
TCTCC	0.7112	0.3773
TTCCT	0.7021	0.4464
CTTCC	0.6997	0.4533
TCCTC	0.6966	0.3538
CCTCC	0.6786	0.3375
TTCCC	0.6654	0.4426
CTCCC	0.6587	0.3347
TCCCT	0.6472	0.3347

A single descriptor, even though it may be instructive, offers but a limited characterization for a large system. Often additional descriptors are needed. They can be constructed by considering the so-called "higher order" **D/D** matrices.[23] These matrices are obtained by taking the powers of the quotients of two distances, rather than just using the quotients of the distances themselves. As a result we can derive for a *geometrical* (graphical-spatial) representation of DNA an *algebraic* characterization based on set of invariants, obtained by calculating the leading eigenvalue of the set of "higher order" matrices ^{n}**D**/^{n}**D**. We will continue to use simplified notation **D/D** even though the **D** in the numerator stands for the Euclidean distances and the **D** in the denominator stands for graph theoretical distances.

D/D MATRICES FOR DNA

The Euclidean distance between bases in a 3-D graphical model of DNA are obtained from the 3-D coordinates of the nucleic bases listed in Table 3 using $\{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\}^{1/2}$, where x_i, y_i, z_i and x_j, y_j, z_j are the Cartesian coordinates of the points considered. To obtain the **D/D** matrix first we have to normalize the distance scale so that the Euclidean distance between adjacent vertices equals 1, not $\sqrt{3}$ (as a result of taking the side of cube to be equal 1). Then we have to divide each Euclidean distance with the number of bonds separating the two vertices to obtain the desired quotient of the two distances. In Table 5 we illustrate a part of the **D/D** matrix (corresponding to nine initial bases of DNA primary sequences of exon 1 of human β gene). The numerator combined with factor $1/\sqrt{3}$ gives the Euclidean distance between vertices i, j when the separation between adjacent bases is assigned distance 1, and

Table 5. Portion of the **D/D** Matrix for the First Exon of DNA of Table 1

0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$4/4\sqrt{3}$	$\sqrt{27}/5\sqrt{3}$	$\sqrt{8}/6\sqrt{3}$	$\sqrt{11}/7\sqrt{3}$	$\sqrt{8}/8\sqrt{3}$
	0	1	$\sqrt{12}/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$\sqrt{24}/4\sqrt{3}$	$\sqrt{19}/5\sqrt{3}$	$\sqrt{12}/6\sqrt{3}$	$\sqrt{11}/7\sqrt{3}$
		0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$\sqrt{8}/4\sqrt{3}$	$\sqrt{3}/5\sqrt{3}$	$2/6\sqrt{3}$
			0	1	$2/2\sqrt{3}$	$\sqrt{3}/3\sqrt{3}$	0	$\sqrt{3}/5\sqrt{3}$
				0	1	$2/2\sqrt{3}$	$\sqrt{3}/3\sqrt{3}$	$\sqrt{8}/4\sqrt{3}$
					0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$
						0	1	$2/2\sqrt{3}$
							0	1
								0

Table 6. Numerical Values for the Initial Portion of **D/D** Matrix and "Higher Order" **D/D** Matrices[a]

0	1	0.57735 0.33333 0.11111 0.12345 1.524−4	0.63828 0.40741 0.16598 0.02755 7.590−4	0.57735 0.33333 0.11111 0.12345 1.524−4	0.60000 0.36000 0.12960 0.01680 2.821−4	0.27217 0.07407 0.00549 3.011−5 9.064−10	0.27355 0.07483 0.00560 3.135−5 9.831−10	0.20412 0.04167 0.00174 3.014−6 9.085−12
	0	1	1	0.63828 0.40741 0.16598 0.02755 7.590−4	0.70711 0.50000 0.25000 0.06250 0.00391	0.50332 0.25333 0.06418 0.00412 1.696−5	0.33333 0.11111 0.12345 1.524−4 2.323−8	0.27355 0.07483 0.00560 3.135−5 9.831−10
		0	1	0.57735 0.33333 0.11111 0.12345 1.524−4	0.63828 0.40741 0.16598 0.02755 7.590−4	0.40825 0.16667 0.02778 7.716−4 5.954−7	0.20000 0.04000 0.00160 2.560−6 6.554−12	0.19245 0.03704 0.00137 1.882−6 3.541−12
			0	1	0.57735 0.33333 0.11111 0.12345 1.524−4	0.33333 0.11111 0.12345 1.524−4 2.323−8	0	0.20000 0.04000 0.00160 2.560−6 6.554−12
				0	1	0.57735 0.33333 0.11111 0.12345 1.524−4	0.33333 0.11111 0.12345 1.524−4 2.323−8	0.40825 0.16667 0.02778 7.716−4 5.954−7
					0	1	0.57735 0.33333 0.11111 0.12345 1.524−4	0.63828 0.40741 0.16598 0.02755 7.590−4
						0	1	0.57735 0.33333 0.11111 0.12345 1.524−4
							0	1
								0

[a] The first row is each box is the numerical value of **D/D** element, while the successive rows correspond to $^{2}\mathbf{D}/^{2}\mathbf{D}$, $^{4}\mathbf{D}/^{4}\mathbf{D}$, $^{8}\mathbf{D}/^{8}\mathbf{D}$, and $^{16}\mathbf{D}/^{16}\mathbf{D}$, respectively.

the denominator is the graph theoretical distance between the same two vertices.

The "higher order" **D/D** matrices are constructed by raising the elements of the **D/D** matrix (Table 5) to an ever increasing power. In Table 6 we show the corresponding entries of the higher order **D/D** matrices which are grouped into a single matrix where each row gives the numerical values corresponding to matrix elements of **D/D**, $^{2}\mathbf{D}/^{2}\mathbf{D}$, $^{4}\mathbf{D}/^{4}\mathbf{D}$, $^{8}\mathbf{D}/^{8}\mathbf{D}$, and $^{16}\mathbf{D}/^{16}\mathbf{D}$. As we can see all matrix elements that are smaller than one decrease as the exponents of the power increase. If one continues to raise exponents to even higher powers all the elements of $^{n}\mathbf{D}/^{n}\mathbf{D}$ matrix that are different from one would soon become very small and could be neglected. Hence, in the limit as $n \rightarrow \infty$ they are zero, and the resulting **D/D** matrix reduces to a binary matrix. In Table 7 we show the initial part of the limiting binary matrix $^{\infty}\mathbf{D}/^{\infty}\mathbf{D}$ for the first exon of DNA of Table 1 again displaying only a 9 × 9 section. As we can see, all the elements above the main diagonal of the limiting matrix corresponding to adjacent sites in the DNA chain are necessarily equal to 1.

Table 7. Initial Portion of the Limiting (Symmetrical) Matrix of $^{n}\mathbf{D}/^{n}\mathbf{D}$ Matrix Truncated at n = 16[a]

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1										
2	1	0	1	1								
3		1	0	1								
4		1	1	0	1							
5				1	0	1						
6					1	0	1					
7						1	0	1				
8							1	0	1			
9								1	0	1	1	
10									1	0	1	
11									1	1	0	1
12											1	0

[a] Only zeros at the diagonal position are shown.

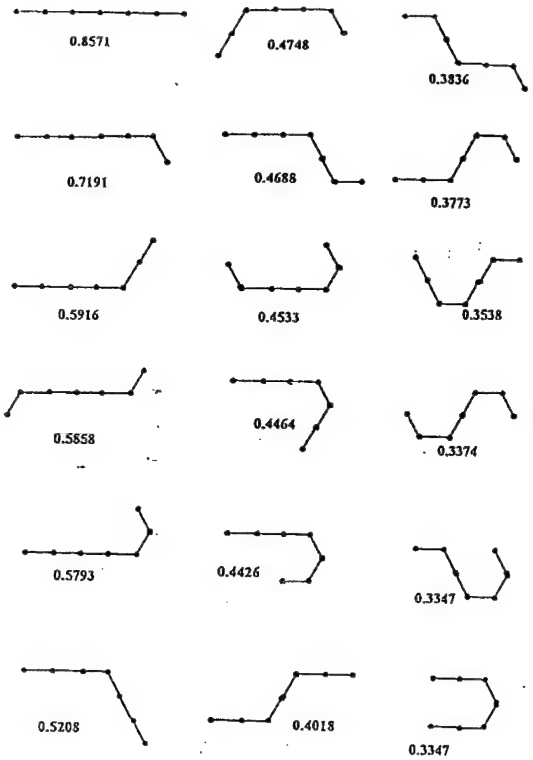


Figure 6. Conformations of line-adjacency graphs of eight-atom chains embedded on a graphite lattice ordered according to decreasing values of leading eigenvalue of line adjacency matrix. The order of isomers in Figure 5 and this figure is different.

However, entry 1 appears in addition at all sites associated with a repetition of the same nucleic base in the primary DNA sequence. For the first exon of Table 1 this happens at sites 3, 4 and 9, 10, and so on. When constructing the 3-D graphical model at these sites we continue to move in the *same* direction, and the corresponding segment of the 3-D graphical model forms a *line* segment. Hence, the elements of the limiting matrix indicate the so-called "line adjacency". The limiting matrix, referred to as the "line adjacency matrix",[24] is known in Graph Theory as the adjacency matrix of the Menger graph of a configuration.[25] For graphs of Figure 5 we show the corresponding Menger graphs. Their "line adjacency" matrix represents the limiting $^{\infty}\mathbf{D}/^{\infty}\mathbf{D}$ matrices. They are also embedded in a plane because they have been derived from already embedded graphs.

A comparison of Figures 5 and 6 shows that line adjacency matrix carries *different* information than the **D/D** matrices from which it was algebraically constructed. The graphs in Figure 5 are ordered according to descending magnitudes of the normalized leading eigenvalue of the adjacency matrix, and the graphs in Figure 6 are ordered according to the leading eigenvalue of the limiting matrix. The resulting order is *different* from the order induced by the leading eigenvalue of **D/D** matrix. The leading eigenvalue of the limiting matrix can be viewed as an index of flexibility (or stiffness) of a structure, at least in some special cases.[24] Apparently structures with longer "line" segments have larger λ_1 or λ_1/n. When this is "translated" to the graphical representation relating to DNA sequences, the occurrence of "straight" segments corresponds to recurrence of the same base in a sequence repeatedly. Hence, DNA sequences with a larger number of repeating bases and longer such repeating segments will have a larger leading eigenvalue of the limiting binary matrix $^{\infty}\mathbf{D}/^{\infty}\mathbf{D}$.

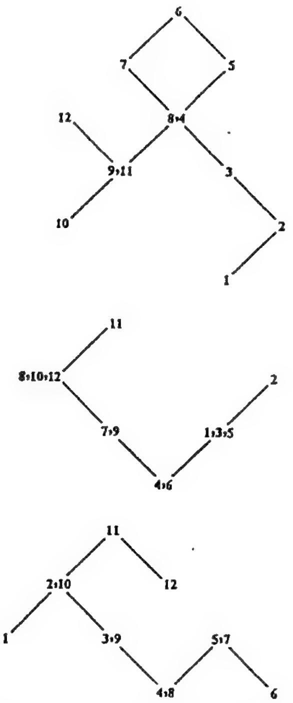


Figure 7. Projection of a portion of 3-D graphical representation of DNA of Figure 4 on the (x, y), (x, z), and (y, z) coordinate planes.

PROJECTIONS OF 3-D SPATIAL SEQUENCE REPRESENTATION

Spatial curves can be projected on coordinate planes (x, y), (x, z) or (y, z), or any plane, for that matter. The projections of 3-D spatial curves on each of the three coordinate planes is quite simple when coordinates of all the points are known. All that is needed is to ignore the coordinate perpendicular to the plane of the projection. Hence, for the first nucleic base of Table 1, A, with spatial coordinates $(+1, -1, -1)$ we have for the projection on the x, y plane $x = 1$ and $y = -1$. For the projection of the same base on the x, z plane we have $x = 1$ and $z = -1$, while for the projection of the first nucleic base on the y, z plane we obtain $y = -1$ and $z = -1$. Hence, the projection coordinates can be read directly from Table 2 by ignoring one column, depending on the projection considered. In Figure 7 we show the three projections for the first 12 bases of exon of DNA of Table 1. It is interesting to observe that projection of the spatial 3-D representation of DNA on the (x, y) coordinate plane is identical with the 2-D graphical representation of Nandy[26,27] already depicted at the top of Figure 2. Hence, our 3-D visual representation of DNA contains automatically the 2-D graphical representation of Nandy as one of its projections. This, however, is not surprising, because if we project the four vertices of the tetrahedron having the coordinates $(+1, -1, -1)$, $(-1, +1, -1)$, $(-1, -1, +1)$, $(+1, +1, +1)$ on the (x, y) plane we obtain points $(+1, -1)$, $(-1, +1)$, $(-1, -1)$, $(+1, +1)$. The first set of points is associated with directions for A, G, C, T in 3-D as outlined in this paper, and the second set of points is associated with directions for A, G, C, T in 2-D that coincides with that of Nandy if we rotate the coordinate system by −135°.

Similarly we find that the projection of the spatial 3-D representation of DNA on the (x, z) coordinate plane is identical with the 2-D graphical representation of Leong and Morgenthaler.[14] Hence, our 3-D visual representation of DNA contains alternative 2-D graphical representations as its projections. We may add that there is third yet the projection of 3-D graphical representation of DNA, the projection on the plane (y, z), that corresponds to the assignment of the four directions defined by the positive and the negative x and y coordinate axes to the four nucleic bases so that A and T are associated with the x-axis and C and G with the y axis. As we see frem Figure 7 this projection differs from those of Nandy, Leong, and Morgenthaler and may have its own merits. Finally, we should add that one can consider projections of 3-D graphical curves of DNA on planes other than coordinate planes. While projections offer convenience of 2-D representation, all these projections are associated with some loss of information associated with the projection process.

Although the three projection paths of the 3-D representation of DNA are different, their limiting matrices are identical. This can be understood, because the form of the limiting matrix depends only on the repetition of same nucleic base in the primary sequence of DNA and that is independent of graphical representation of DNA and the projection process.

MATRIX INVARIANTS OF DNA

The search for a matrix representation of DNA primary sequence was motivated by desire to have numerical descriptors for DNA that are sequence invariants. Numerical characterization of DNA primary sequences will make comparisons of different DNA sequences much simpler than comparison based on alphabet symbols or the corresponding codes. Moreover, it will lead to quantitative measure of similarity and may open a novel method of characterizations for the same set of sequences. Matrices not only offer various inherent invariants as a tool for such comparisons but also allow one to consider modifications of matrix elements and in this way may further enrich the tool for comparative study of DNA. In this report we will continue to confine our attention to **D/D** matrix of DNA, but it will be clear that the outlined schemes are equally valid not only for the "higher order" **D/D** matrices but also for other matrices that one can associate with DNA.

Table 8. The Upper Bounds, the Lower Bounds, the Leading Eigenvalue, and the Average Row Sums for Truncated Matrices of DNA

	row sum max	λ_1	row sum min	row sum average
1	0	0	0	0
2	1	1	1	1
3	2	1.732051	1.57735	1.718233
4	3	2.629245	2.21563	2.607815
5	3.63828	3.238402	2.79298	3.203444
6	4.34539	3.869193	3.39298	3.843783
7	4.84871	4.242930	3.09442	4.178791
8	5.18204	4.455833	2.71756	4.335833
9	5.45559	4.737987	3.49400	4.508241

Table 9. Average Matrix Element as a Function of Gradually Truncated **D/D** Matrix

	x, y, z	x, y	x, z	z, y
1	0	0	0	0
2	0.86603	0.70711	0.70711	0.70711
3	1.21424	1.07298	0.62854	1.07298
4	1.74711	1.52917	1.06066	1.52917
5	2.00204	1.82479	0.90510	1.82479
6	2.34274	2.16431	1.02138	2.16431
7	2.35303	2.25833	1.23982	2.07952
8	2.23832	2.11133	1.21440	1.97442
9	2.25630	2.13265	1.24376	1.89102
10	2.46497	2.24965	1.54132	1.94565
11	2.51032	2.19350	1.71264	2.03576
12	2.55077	2.23357	1.80319	2.00313
13	2.47111	2.15924	1.75222	1.92259
14	2.51231	2.20976	1.79277	1.89779
15	2.50930	2.12249	1.84061	1.92616
16	2.63879	2.14294	2.01366	2.04107

Among numerous matrix (and graph) invariants we will consider first the average matrix element, which in the case of the graph theoretical distance matrix, except for normalization, is related to the Wiener number, a well-known graph theoretical invariant.[28,29] Alternatively one can consider the average row sum, which differs from the average matrix element and the Wiener number again only by normalization factor. The average row sum has an advantage, particularly when the individual row sums do not differ widely, because it may suggest an approximate value for the leading eigenvalue of the matrix. According to the Frobenius−Perron theorem of linear algebra the largest and the smallest row sums represent the upper and the lower bounds, respectively, for the leading eigenvalue (λ_1) of a symmetric matrix.[30] In Table 8 we have listed the upper bounds, the lower bounds, and the leading eigenvalue for truncated sequence of DNA for $n = 1$ to $n = 9$. Observe how closely the average row sum (given in the last column) approximates the leading eigenvalue, particularly for shorter segments of the matrix.

The leading eigenvalue of a matrix is an important matrix invariant. We have already mentioned that λ_1/n of the **D/D** matrix is an index of the folding of a structure, and λ_1/n of the limiting matrix can be viewed as an index of the flexibility of a system. Similarly, the λ_1 of the adjacency matrix and λ_1 of the path matrix represent alternative indices of (molecular) branching,[31,32] while λ_1 of the **D/DD** matrix, where **DD** represents the detour matrix,[33,34] is an index of the cyclicity of a system.[35,36] The average row sum may give a similar insight into a system as the leading eigenvalue. The average row sum, however, can be easily computed, while computation of eigenvalues of large matrices is more involved, and, of course, the DNA sequences could be very long. For example, the 1424 bases of Table 1, of which we considered the first exon only (92 bases), are a part of 73 326 base pairs.[37]

The average row sum, and also the average matrix element of a **D/D** matrix, will depend on the size of the matrix as is seen from Table 9 where under the heading x, y, z we have listed the average matrix element as a function of n, the size of the matrix at truncation of DNA sequence. The same was true for the leading eigenvalue of the truncated DNA sequences (Table 8).

The dramatic condensation of data illustrated above may be excessive for some more ambitious comparisons of DNA sequences. In such cases, one can, in addition to **D/D** matrix, also consider the leading eigenvalue or the average element

Table 10. Leading Eigenvalue of the **D/D** Matrix and Higher Order **D/D** Matrices for $n = 2$ to $n = 20$ Showing the Convergence for λ_1 and the Limit for $n \to \infty$

power	λ_1	power	λ_1
1	4.73797	12	2.35418
2	3.54855	13	2.35143
3	2.99558	14	2.34966
4	2.71223	15	2.34851
5	2.55903	16	2.34777
6	2.47313	17	2.34729
7	2.42349	18	2.34696
8	2.39409	19	2.34675
9	2.37629	20	2.34661
10	2.36537	limit	2.34631654447882
11	2.35850		

of ^{2}**D**/^{2}**D** matrix, of ^{3}**D**/^{3}**D** matrix, and so on. A dozen ^{n}**D**/^{n}**D** matrices can in this way offer a sufficient number of invariants for more extensive comparisons of DNA sequences. In Table 10 we report the leading eigenvalue for a 9 × 9 ^{n}**D**/^{n}**D** matrices for $n = 1$ to $n = 20$, which illustrate the "profile," the sequence of descriptors, for the particular fragment of DNA. As n increases the value of the leading eigenvalue λ_1 converges to a limiting value. The limit can be easily computed as it represents the leading eigenvalue of the binary matrix of the same size (here 9 × 9). Using so constructed "profiles" the calculation of the similarities of DNA sequences is transformed into a calculation of similarities of the corresponding numerical sequences of DNA descriptors, the task which is not computer intensive if compared to the similar studies using alignment methodologies. Of course, it yet remains to be investigated which set of invariants may offer optimal characterization for DNA comparisons and how sensitive are such "profiles" to minor changes in DNA composition. In a recent study in which the DNA sequence was characterized by average distances between various nucleic acid bases it was shown that the "distance profiles", constructed analogously to the here reported "leading eigenvalue profile", is very sensitive already when a single nucleic base has been changed (i.e., the case of mutation).[41]

CONCLUDING REMARKS

In this article we (1) outlined a construction of a 3-D "graphical" representation of DNA primary sequences, illustrated on a portion of the human β globin gene; (2) described a particular scheme that allows 3-D spatial representation of DNA to be transformed into a numerical matrix representation; (3) illustrated derivation of a set of matrix invariants from the matrix representation of DNA; and (4) suggested a relative simple data reduction based on statistical analysis of generated DNA matrix invariants. Each of the four contributions, in our view, not only will facilitate comparative studies of DNA but also open possibilities for further developments of condensation of primary DNA sequence information. The outlined 3-D representation, for example, can be modified by use of the sequential labels as the fourth coordinate in order to avoid 3-D spatial curves overlap itself. The numerical matrix characterization offers many alternatives, from the use of different distance measures to the use of different matrix forms. In addition to the possibility of selecting matrix invariants, which is almost unlimited, we have the possibility of selecting different matrices to start the process of condensation of data. Hence, we anticipate here an expansion, if not explosion, of alternatives that may parallel the expansion of the topological indices proposed for the characterization of molecular structure–property-activity relationships and introduction of novel matrices for chemical graphs. The most significant aspect considered in this contribution may turn out to be the data reduction step when a large number of input data are condensed into a substantially smaller set of derived parameters. This important aspect of DNA data analysis has only recently received some attention,[38–40] but, in view of the exponential growth of the automated DNA sequencing techniques, the problem of digesting novel information, no doubt, will require novel ideas that go beyond just listings of nucleic bases of a primary sequence. The construction of sequence "profiles", illustrated in this report, may be one way of data reduction, in addition to the recently proposed grouping of data for different nucleic acids separately, which allow large ($n \times n$) matrices (where n can run into the hundreds or the thousands) to be condensed to small (4 × 4) matrices where the rows and the columns are associated with the four nucleic bases A, G, C, and T. Needless to say that the outlined approach is suitable for characterization of local fragments of DNA, which is precisely how one may look on the truncated DNA fragment considered in this work.

ACKNOWLEDGMENT

We thank Professor Ch. Rücker (Freiburg, Germany) for critical reading of the manuscript and numerous suggestions that lead to improved presentation of the material.

REFERENCES AND NOTES

(1) Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernet. Control Theor.* **1966**, *10*, 707–710.

(2) Sankoff, D. Matching sequences under deletion-insertion constraints. *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *68*, 4–6.

(3) Kruskal, J. B. An overview of sequence comparison. In *Time wraps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*; Sankoff, D., Kruskal, J. B., Eds.; Addison-Wesley: London, 1983; pp 1–40.

(4) Waterman, M. S. General methods of sequence comparison. *Bull. Math. Biol.* **1984**, *46*, 473–500.

(5) Smith, T. F.; Waterman, M. S. Comparison of biosequences. *Adv. Appl. Math.* **1981**, *2*, 482–489.

(6) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197.

(7) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444–2448.

(8) Jerman-Blazic, B.; Fabič, I.; Randić, M. Comparison of sequences as a method for evaluation of the molecular similarity. *J. Comput. Chem.* **1986**, *7*, 176–188.

(9) Jerman-Blažič, B.; Fabič, I.; Randić, M. Application of string comparison techniques in QSAR Studies. In *QSAR in Drug Design and Toxicology;* Hadzi, D., Jerman-Blažič, B., Eds.; Elsevier Sci. Publ.: Amsterdam, The Netherlands, 1987; pp 52–54.

(10) Randić, M. *Topological indices, The Encyclopedia of Computational Chemistry;* Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley and Sons: Chichester, 1998; pp 3018–3032.

(11) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672–687.

(12) Randić, M.; Basak, S. C. Variable molecular descriptors. In *Some Aspects of Mathematical Chemistry;* Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; to be published by Visva Bharati University, Santiniketan, West Bengal, India

(13) Roy, A.; Raychaudhury, C.; Nandy, A. A novel techniques of graphical representation and analysis of DNA sequences – A Review. *J. Biosci.* **1998**, *23*, 55.

(14) Leong, P. M.; Mogenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *12*, 503–511.

(15) Hamori, E. Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques.* **1989**, *7*, 710–720.

(16) Nandy, A. New graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309.

(17) Randić, M.; Nandy, A.; Basak, S. C. On numerical characterization of DNA primary sequences. *J. Math. Chem.*, submitted for publication.

(18) Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.

(19) Harary, F. *Graph Theory;* Addison-Wesley: Reading, MA, 1969.

(20) Buckley, F.; Harary, F. *Distance in Graphs;* Addison-Wesley: Reading, MA, 1990.

(21) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277.

(22) Randić, M.; Razinger, M. *On characterization of 3D molecular structure,* in: *From Chemical Topology to Three-Dimensional Geometry;* Balaban, A. T., Ed.; Plenum Press: New York, 1997.

(23) Randić, M.; Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017–1026.

(24) Randić, M.; Vračko, M.; Novič, M. *Eigenvalues as molecular descriptors,* in: *QSAR/QSPR by Molecular Descriptors;* Diudea, M. V., Ed.; Nova Publ.: in press.

(25) Coxeter, H. S. M. *Bull. Am. Math. Soc.* **1950**, *56*, 413.

(26) Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci.* **1996**, *70*, 661–668.

(27) Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for maromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.

(28) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(29) Trinajstić, N. *Chemical Graph Theory,* 2nd ed.; CRC Press: Boca Raton, Fl, 1992.

(30) Gantmacher, F. *Theory of Matrices;* Chelsea Publ: New York, 1959; Vol. II, Chapter 13.

(31) Randić, M. On molecular branching. *Acta Chim. Slovenica* **1997**, *44*, 57–77.

(32) Randić, M. J. On structural ordering and branching of acyclic saturated hydrocarbons. *Math. Chem.* **1998**, *24*, 345–358.

(33) Amić, D.; Trinajstić, N. On the detour matrix *Croat. Chem. Acta* **1995**, *68*, 53–62.

(34) Trinajstić, N.; Nikolić, S.; Lučuć, B.; Amić, D.; Mihalić, Z. The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 631.

(35) Randić, M. J. On characterization of cyclic structures. *Chem. Inf. Comput. Sci.* **1997**, *37*, 1063.

(36) Pisanski, T.; Plavšić, D.; Randić, M. On numerical characterization of cyclicity *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 520–523.

(37) EMBL Nucleic Bases Sequence Database (rel. 31) ID HSHBB Accession number V01317.

(38) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50–56.

(39) Randić, M.; Vračko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606.

(40) Randić, M. On characterization of DNA primary sequences by a condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29–34.

(41) Randić, M.; Basak, S. C. Characterization of DNA based on the average distances between the nucleic acid bases. *J. Chem. Inf. Comput. Sci.*, submitted for publication.

CI000034Q

APPENDIX 1.10 QSPR modeling: Graph connectivity indices versus line graph connectivity indices

QSPR Modeling: Graph Connectivity Indices versus Line Graph Connectivity Indices[†]

Subhash C. Basak, Sonja Nikolić,[‡] and Nenad Trinajstić*,[‡]

Natural Resources Research Institute, University of Minnesota, Duluth, Minnesota 55811

Dragan Amić and Drago Bešlo

Faculty of Agriculture, The Josip Juraj Strossmayer University, HR-31001 Osijek, The Republic of Croatia

Received September 9, 1999

Five QSPR models of alkanes were reinvestigated. Properties considered were molecular surface-dependent properties (boiling points and gas chromatographic retention indices) and molecular volume-dependent properties (molar volumes and molar refractions). The vertex- and edge-connectivity indices were used as structural parameters. In each studied case we computed connectivity indices of alkane trees and alkane line graphs and searched for the optimum exponent. Models based on indices with an optimum exponent and on the standard value of the exponent were compared. Thus, for each property we generated six QSPR models (four for alkane trees and two for the corresponding line graphs). In all studied cases QSPR models based on connectivity indices with optimum exponents have better statistical characteristics than the models based on connectivity indices with the standard value of the exponent. The comparison between models based on vertex- and edge-connectivity indices gave in two cases (molar volumes and molar refractions) better models based on edge-connectivity indices and in three cases (boiling points for octanes and nonanes and gas chromatographic retention indices) better models based on vertex-connectivity indices. Thus, it appears that the edge-connectivity index is more appropriate to be used in the structure–molecular volume properties modeling and the vertex-connectivity index in the structure–molecular surface properties modeling. The use of line graphs did not improve the predictive power of the connectivity indices. Only in one case (boiling points of nonanes) a better model was obtained with the use of line graphs.

INTRODUCTION

This study was motivated by two recent papers. In one Estrada and Rodriguez[1] have shown that the edge-connectivity index produced the best single-variable QSPR models for five out of seven physicochemical properties of octanes. In another Gutman et al.[2] have reported that the use of line graphs, in some cases, significantly improves the predictive power of topological indices. We decided to test both of these results by using them to reinvestigate several QSPR models from the literature. We also decided to test further the result that in many cases the optimum exponent of the vertex- and edge-connectivity indices is not -0.5.[3] Since we believe, along with many others,[4] that the QSPR modeling will become the tool of choice for many chemists-at-large in times to come, it seems to us worthwhile to search for the most reliable framework to carry out this kind of modeling. The present study is an attempt in this direction. It should also be noted that throughout this paper we will use the chemical graph theoretical concepts and language[5] only to simplify the analysis.

Recently, line graphs have been increasingly used in structure–property modeling,[2,6–11] although they may be traced back to van't Hoff, who used the line graphs of the structural formulas for representing simple organic compounds. Line graphs are described in a monograph on chemical graph theory[12] and under the name bond graphs were used in deriving the molecular complexity indices.[13] The line graph L(G) = L of graph G is a graph derived from G in such a way that the edges in G are replaced by vertexes in L. Two vertexes in L are adjacent if the corresponding two edges in G are incident, that is, have a vertex in common. The construction of a line graph from a tree is shown in Figure 1.

The line graph L is usually a more complex structure than the corresponding graph G. Only in the case of unbranched cycloalkanes, represented by cycles, L and G coincide because in cycles the number of vertexes V and the number of edges E are identical. For n-alkanes, represented by the hydrogen-depleted chains, L is less complex than G because it has one less vertex than G, since in chains $E = V - 1$.

The numbers of vertexes V and edges E of the line graph L and the corresponding graph G are related by

$$V(\mathrm{L}) = E(\mathrm{G}) \tag{1}$$

$$E(\mathrm{L}) = (1/2)\sum_i d_i^2(\mathrm{G}) - E(\mathrm{G}) \tag{2}$$

where d_i ($i = 1, 2, ..., V$) are degrees of vertexes in G. These relations can be easily confirmed by inspecting G and L depicted in Figure 1.

Using the equation

$$\sum_i d_i^2(\mathrm{G}) = M_1 \tag{3}$$

where M_1 is called[14–16] the first Zagreb-group index,[17,18] and

[†] Reported in part at the One Day Symposium on Applied Mathematical Chemistry, held on May 3, 1999, at the Natural Resources Research Institute, University of Minnesota, Duluth.

[‡] On leave of absence from The Rugjer Bošković Institute, HR-10001 Zagreb, The Republic of Croatia.

10.1021/ci990119v CCC: $19.00 © 2000 American Chemical Society
Published on Web 04/07/2000

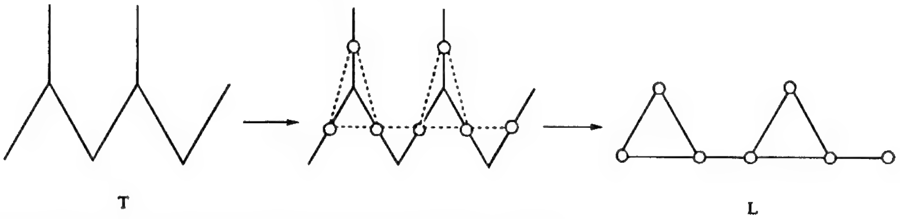


Figure 1. Construction of line graph L from tree T depicting 2,4-dimethylhexane.

introducing (3) into (2), we obtain the expression

$$E(\mathrm{L}) = (1/2)M_1 - E(\mathrm{G}) \qquad (4)$$

Gutman and Estrada derived the same expression,[19] but the factor 1/2 is missing in their expression. From (4) follows an amusing result that the M_1 index of the graph is simply equal to twice the number of vertexes and edges in the line graph:

$$M_1 = 2[E(\mathrm{L}) + E(\mathrm{G})] = 2[E(\mathrm{L}) + V(\mathrm{L})] \qquad (5)$$

SIMPLE MODIFICATION OF THE VALENCE VERTEX- AND EDGE-CONNECTIVITY INDICES

Vertex-Connectivity Index. The standard definition of the vertex-connectivity index is[20]

$$\chi = \sum_{\text{edges}} [d(v_i)\, d(v_j)]^{-0.5} \qquad (6)$$

where $d(v_i)$ is the degree of the vertex v_i and $[d(v_i)\, d(v_j)]^{-0.5}$ may be considered as the weight of the $i-j$ edge.[21] The summation in (6) goes over all edges. The vertex degree $d(v_i)$ is equal to the number of vertexes adjacent to vertex i in a graph G. Any two vertexes in G are adjacent if there are edges connecting them.

Equation 6 is open to modification because the choice of edge weights $[d(v_i)\, d(v_j)]^{-0.5}$ was based on one possible solution to the inequalities based on ordering graphs.[20] There are also other choices of weights possible. Hence, the quantity $[d(v_i)\, d(v_j)]^{-0.5}$ can be replaced by $[d(v_i)\, d(v_j)]^k$, where k is a variable exponent that can be varied in any desired range of values, and (6) becomes[3]

$$\chi = \sum_{\text{edges}} [d(v_i)\, d(v_j)]^k \qquad k \neq 0 \qquad (7)$$

Edge-Connectivity Index. The standard definition of the edge-connectivity index is similar to the definition of the vertex-connectivity index, the only change being in using the edge degrees $d(e_i)$ instead of vertex degrees $d(v_i)$:[22]

$$\epsilon = \sum_{\text{adjacent edges}} [d(e_i)\, d(e_j)]^{-0.5} \qquad (8)$$

The edge degree $d(e_i)$ is equal to the number of edges adjacent to edge i in a graph G. Any two edges in G are adjacent if they meet at the same vertex. Because every edge in G connects two vertexes, the edge degree $d(e)$ can be expressed in terms of their degrees as follows:[22]

$$d(e) = d(v_i) + d(v_j) - 2 \qquad (9)$$

This expression can be used to assign the degrees of edges in G. In Figure 2 we give the vertex and edge degrees in tree T and the corresponding line graph L depicted in Figure 1.

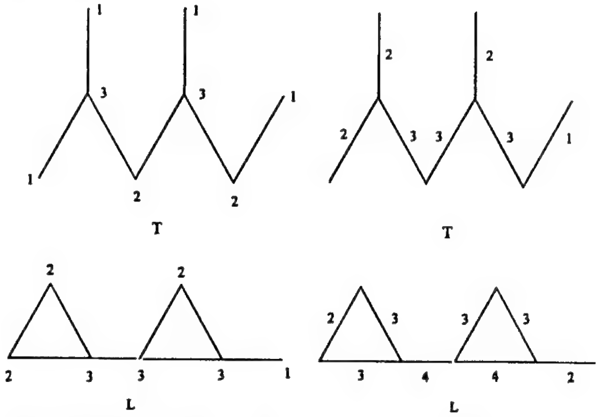


Figure 2. Vertex degrees (digits at each vertex) and edge degrees (digits at each edge) in tree T and the corresponding line graph L from Figure 1.

A simple way to assign the degrees to edges in graph G or its line graph L is to count all adjacent bonds of a bond for which we wish to determine the edge degree. This procedure is illustrated in Figure 3.

Equation 8 can also be modified because the quantity $[d(e_i)\, d(e_j)]^{-0.5}$ was the result of mimicking the original definition of Randić for the vertex-connectivity index.[20] Consequently, $[d(e_i)\, d(e_j)]^{-0.5}$ can be replaced by $[d(e_i)\, d(e_j)]^k$, where k is a variable exponent that can be varied in any desired range of values. Thus, (8) converts into the following equation:

$$\epsilon = \sum_{\text{adjacent edges}} [d(e_i)\, d(e_j)]^k \qquad k \neq 0 \qquad (10)$$

At this point it should also be noted that the edge-adjacency matrix[23] of the graph G, $^{\mathrm{E}}\mathrm{A}(\mathrm{G})$, is identical to the vertex-adjacency matrix[23] of the line graph L of G, $^{\mathrm{V}}\mathrm{A}(\mathrm{L})$:

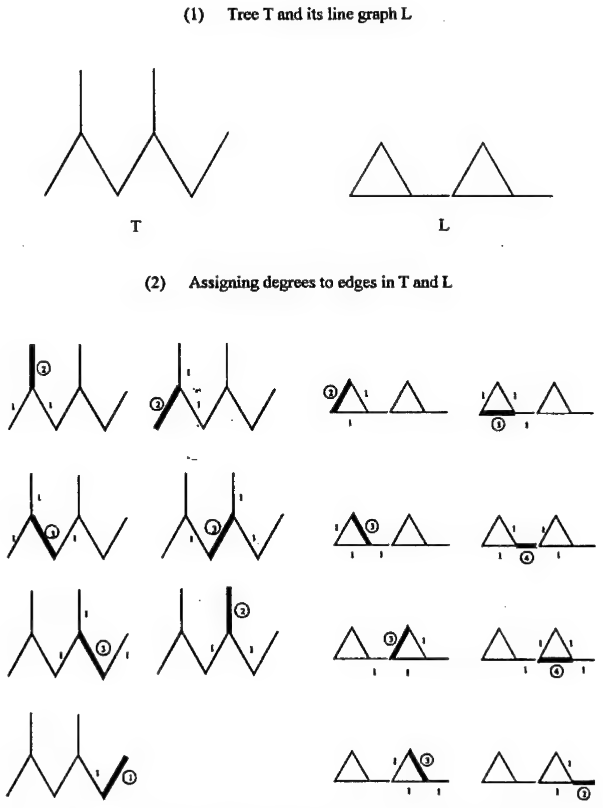
$$^{\mathrm{E}}\mathrm{A}(\mathrm{G}) = {}^{\mathrm{V}}\mathrm{A}(\mathrm{L}) \qquad (11)$$

This must be so because the edge degrees in G are identical to the vertex degrees in the corresponding line graph L (see Figure 2). The consequence of (11) is that the edge-connectivity index of G is identical to the vertex-connectivity index of the corresponding line graph L:[19]

$$\epsilon(\mathrm{G}) = \chi(\mathrm{L}) \qquad (12)$$

RESULTS AND DISCUSSION

We studied five structure–property models that were already reported in the literature. This was done on purpose because our aim was to compare the performance of the obtained models with those already published. The properties considered were boiling points of octanes and nonanes and


(1) Tree T and its line graph L

(2) Assigning degrees to edges in T and L

Figure 3. A simple procedure for assigning the degrees to edges in tree T and the related line graph L.

molar volumes, molar refractions, and retention indices of alkanes. Boiling points and retention indices are typical "surface"-dependent properties, while molar volumes and molar refractions are "molecular volume"-dependent properties. In all cases molecules were depicted as graphs and corresponding line graphs. The standard deviation S was used as a criterion for the comparison of the models. The optimum parameter k in (7) and (10) was determined using the procedure described in our earlier report;[3] that is, the parameter k was taken to be optimum when the value of S reached a minimum.

Boiling Points of 18 Octanes. We first considered structure–boiling point models for isomeric octanes, on the basis of their vertex-connectivity indices computed for octane trees. The best model was obtained for $k = -1.15$. The regression equation is given by

$$\text{bp} = 65.14(\pm 7.29) + 28.87(\pm 4.31)\chi^{[-1.15]} \quad (13)$$
$$n = 18 \quad R = 0.859 \quad S = 3.24 \quad F = 45$$

where bp is the normal boiling point, R the correlation coefficient, S the standard deviation, F the Fisher ratio, and $\chi^{[-1.15]}$ a short-hand notation for the vertex-connectivity index computed using the value of -1.15 for the exponent in (7). The notation $\chi^{[k]}$ will be used throughout this paper. The improvement over the model based on $k = -0.5$ is rather slight:

$$\text{bp} = 3.14(\pm 19.23) + 30.33(\pm 5.27)\chi^{[-0.50]} \quad (14)$$
$$n = 18 \quad R = 0.821 \quad S = 3.60 \quad F = 33$$

The above models are identical to structure–boiling point models for octanes published elsewhere.[3,24] Randić et al.[25] have also observed that the modified vertex-connectivity index produces better structure–boiling point models of lower (C_2–C_7) alkanes than the standard version of the vertex-connectivity index. However, they have found that the exponent value of -0.33 leads to the best models of three alternatives they considered ($k = -0.5, -0.33, -0.25$).

The same analysis as with the vertex-connectivity index was also carried out with the edge-connectivity index. The best model was obtained for $k = -0.30$. The regression equation is given by

$$\text{bp} = 179.75(\pm 11.12) - 13.66(\pm 2.30)\epsilon^{[-0.30]} \quad (15)$$
$$n = 18 \quad R = 0.830 \quad S = 3.52 \quad F = 35$$

where $\epsilon^{[-0.30]}$ is a short-hand notation for the edge-connectivity index computed using the value of -0.30 for the exponent in (10). The notation $\epsilon^{[k]}$ will be used throughout this paper. The improvement over the model based on $k = -0.5$ is considerable

$$\text{bp} = 162.76(\pm 29.94) - 14.52(\pm 8.85)\epsilon^{[-0.50]} \quad (16)$$
$$n = 18 \quad R = 0.379 \quad S = 5.84 \quad F = 3$$

but the model in (15) is not as good as the model in (13), though it is somewhat better than the model in (14). This result supports the work by Estrada and Rodriguez,[1] because one of the two physicochemical properties of octanes for which the use of the edge-connectivity index did not produce the best single-variable QSPR model was the boiling point, the other being the heat of vaporization. Estrada and Rodriguez pointed out that to describe these properties correctly it is necessary to take into account long-range contributions in the edge-connectivity index.[9] In both these cases better single-variable models were obtained using the Hosoya Z index.[26]

Finally, we considered octane line graphs. Since $\chi^{[k]}(L) = \epsilon^{[k]}(G)$, we derived structure–boiling point models based on the edge-connectivity index $\epsilon^{[k]}(L)$. The best model was obtained for $k = -0.675$. The regression equation is given by

$$\text{bp} = 167.56(\pm 9.03) - 20.17(\pm 3.37)\epsilon^{[-0.675]}(L) \quad (17)$$
$$n = 18 \quad R = 0.831 \quad S = 3.51 \quad F = 36$$

where $\epsilon^{[-0.675]}(L)$ is a short-hand notation for the edge-connectivity index computed for a line graph using the value of -0.675 for the exponent in (10). This notation will be used throughout this paper when the models based on line graphs and edge-connectivity indices are discussed.

The model in (17) is practically the same as the model in (15) on the basis of octane trees and the edge-connectivity index. The improvement over the model based on $k = -0.5$ is visible:

$$\text{bp} = 138.83(\pm 5.80) - 6.11(\pm 1.39)\epsilon^{[-0.50]}(L) \quad (18)$$
$$n = 18 \quad R = 0.740 \quad S = 4.24 \quad F = 19$$

However, this model is much better than the corresponding model in (16) on the basis of octane trees.

Boiling Points of 35 Nonanes. The same kind of analysis as in the case of modeling boiling points of octanes is carried out for nonanes. We first considered structure–boiling point models for isomeric nonanes, on the basis of their vertex-connectivity indices computed for nonane trees. The best model was obtained for $k = -1.25$. The regression equation is given by

$$\mathrm{bp} = 94.23(\pm 6.68) + 25.58(\pm 3.97)\chi^{[-1.25]} \tag{19}$$
$$n = 35 \qquad R = 0.746 \qquad S = 4.13 \qquad F = 41$$

The improvement over the model based on $k = -0.5$ is again rather slight:

$$\mathrm{bp} = 31.47(\pm 19.64) + 25.67(\pm 4.77)\chi^{[-0.50]} \tag{20}$$
$$n = 35 \qquad R = 0.683 \qquad S = 4.53 \qquad F = 29$$

The above models are comparable to the structure–boiling point models for nonanes published elsewhere.[3,27] The same analysis was also carried out with the edge-connectivity index. The best model was obtained for $k = -0.375$. The corresponding regression equation is

$$\mathrm{bp} = 225.36(\pm 16.39) - 18.42(\pm 3.41)\epsilon^{[-0.375]} \tag{21}$$
$$n = 35 \qquad R = 0.685 \qquad S = 4.52 \qquad F = 29$$

This model and the model in (20) are practically the same. However, it is worse than the model in (19). The improvement over the model based on $k = -0.5$ is considerable:

$$\mathrm{bp} = 218.35(\pm 30.12) - 21.31(\pm 7.89)\epsilon^{[-0.50]} \tag{22}$$
$$n = 35 \qquad R = 0.426 \qquad S = 5.61 \qquad F = 7$$

Finally, we considered nonane line graphs. The best model was obtained for $k = -0.70$:

$$\mathrm{bp} = 203.18(\pm 9.60) - 22.96(\pm 3.32)\epsilon^{[-0.70]}(\mathrm{L}) \tag{23}$$
$$n = 35 \qquad R = 0.769 \qquad S = 3.97 \qquad F = 48$$

This model is better than any regarding the relationship between structures and boiling points of nonanes. It represents an improvement over the model based on $k = -0.5$:

$$\mathrm{bp} = 161.56(\pm 5.94) - 22.96(\pm 3.32)\epsilon^{[-0.50]}(\mathrm{L}) \tag{24}$$
$$n = 35 \qquad R = 0.587 \qquad S = 5.03 \qquad F = 17$$

Comparison between this model and the related models based on octane trees shows that the model in (24) is not as good as the model in (20), but better than the model in (22).

In this example, the edge-connectivity index did live up to the expectations based on the work by Gutman et al.:[2] The use of the line graph edge-connectivity index produced for nonanes the best structure–boiling point model. However, the model in (23) is still far from being satisfactory in comparison with models that use several topological indices.[28] For example, the best structure–boiling point model for nonanes with five descriptors has $R = 0.981$ and $S = 0.89$.[29]

Gas Chromatographic Retention Indices of Alkanes. The same methodology as above was applied to the relationship between the structures of alkanes and their gas chromatographic retention indices.[30] We first considered structure–chromatographic retention data correlation for the first 157 alkanes using as the structural parameter the vertex-connectivity index. The best correlation was obtained for $k = -0.325$:

$$\mathrm{RI} = 74.58(\pm 8.48) + 148.14(\pm 1.53)\chi^{[-0.325]} \tag{25}$$
$$n = 157 \qquad R = 0.992 \qquad S = 23.8 \qquad F = 9330$$

where RI stands for the retention indices of alkanes. This model gives a very good agreement between experimental and computed retention indices of alkanes. Retention indices of alkanes cover a range from RI(methane) = 100 to RI-(2,3-dimethylundecane) = 1251.4. In most cases the difference between experimental and computed values is less than 3%.

The model in (25) is only slightly better than the model based on $k = -0.5$:

$$\mathrm{RI} = 64.92(\pm 9.38) + 187.97(\pm 2.13)\chi^{[-0.50]} \tag{26}$$
$$n = 157 \qquad R = 0.990 \qquad S = 26.0 \qquad F = 7801$$

The use of the edge-connectivity index produced poorer models:

$$\mathrm{RI} = 137.98(\pm 13.79) + 200.54(\pm 3.66)\epsilon^{[-0.55]} \tag{27}$$
$$n = 157 \qquad R = 0.975 \qquad S = 41.3 \qquad F = 3008$$

$$\mathrm{RI} = 134.0(\pm 14.55) + 184.89(\pm 3.54)\epsilon^{[-0.50]} \tag{28}$$
$$n = 157 \qquad R = 0.973 \qquad S = 43.2 \qquad F = 2729$$

These two models are comparable, but are much better than models based on alkane line graphs and their edge-connectivity indices:

$$\mathrm{RI} = 206.58(\pm 21.72) + 262.94(\pm 8.30)\epsilon^{[-0.775]}(\mathrm{L}) \tag{29}$$
$$n = 157 \qquad R = 0.931 \qquad S = 68.2 \qquad F = 1003$$

$$\mathrm{RI} = 365.44(\pm 36.63) + 104.00(\pm 7.24)\epsilon^{[-0.50]}(\mathrm{L}) \tag{30}$$
$$n = 157 \qquad R = 0.756 \qquad S = 122.2 \qquad F = 206$$

There are several structure–chromatographic retention index correlations for alkanes available in the literature.[30] Most of them are based on the two-dimensional and three-dimensional Wiener numbers. However, there is also a correlation available based on the vertex-connectivity index with $k = -0.5$ which differs only slightly from (26):[30]

$$\mathrm{RI} = 69.81(\pm 9.31) + 186.93(\pm 2.11)\chi^{[-0.50]} \tag{31}$$
$$n = 157 \qquad R = 0.990 \qquad S = 26.0 \qquad F = 7827$$

The initial work on the structure–chromatographic retention data correlations is due to Randić.[31] The correlations based on the two-dimensional (${}^2W$) and three-dimensional (${}^3W$) Wiener numbers, which are adjusted Walker-type correlations,[32] are not as good as the model in (25):[30]

$$\mathrm{RI} = 171.2(\pm 15.7)\,{}^2W^{0.335(\pm 0.013)} - 48.6(\pm 27.3) \tag{32}$$
$$n = 157 \qquad R = 0.984 \qquad S = 33.0 \qquad F = 2403$$

$$\mathrm{RI} = 170.6(\pm 17.0)\,{}^3W^{0.325(\pm 0.013)} - 31.8(\pm 30.2) \tag{33}$$
$$n = 157 \qquad R = 0.982 \qquad S = 35.6 \qquad F = 2048$$

These models are, however, better than the ones based on edge-connectivity indices computed for either alkane trees or alkane line graphs. The best overall structure−chromatographic retention data correlation is obtained with the vertex-connectivity index with $k = -0.325$ (model in (25)). This is to our knowledge the best structure−chromatographic retention data model of alkanes that exists in the literature.

Molar Volumes of Alkanes. We considered molar volumes of 69 lower alkanes taken from Estrada.[22] We first considered the structure−molar volume relationship using the vertex-connectivity index. The best correlation was obtained for a rather small value of k (−0.07). The regression equation and the statistical parameters for the correlation are:

$$\mathrm{MV} = 55.85(\pm 2.10) + 16.53(\pm 0.32)\chi^{[-0.07]} \qquad (34)$$

$$n = 69 \qquad R = 0.988 \qquad S = 2.73 \qquad F = 2649$$

where MV stands for molar volume. This regression is better as expected than the one based on the standard value of k (−0.50):

$$\mathrm{MV} = 53.07(\pm 4.41) + 29.60(\pm 1.18)\chi^{[-0.50]} \qquad (35)$$

$$n = 69 \qquad R = 0.951 \qquad S = 5.38 \qquad F = 632$$

These models are inferior to those based on the edge-connectivity index. The best structure−molar volume model was obtained for $k = -0.515$:

$$\mathrm{MV} = 57.44(\pm 1.37) + 31.80(\pm 0.41)\epsilon^{[-0.515]} \qquad (36)$$

$$n = 69 \qquad R = 0.995 \qquad S = 1.81 \qquad F = 6094$$

This model is only very slightly better than the model based on the standard value of the exponent k:

$$\mathrm{MV} = 58.23(\pm 1.41) + 30.80(\pm 0.41)\epsilon^{[-0.50]} \qquad (37)$$

$$n = 69 \qquad R = 0.994 \qquad S = 1.88 \qquad F = 5669$$

Equation 37 is different from the corresponding one given by Estrada[22] as (1) in his paper. The difference is caused by the use of erroneous values of the edge-connectivity indices for six alkanes in Table 1 in Estrada's paper. The correct values are (we use the same codes for alkanes as Estrada): (33ME5) −3.1160, (233MMM5) −3.2832, (33ME6) −3.6766, (234MMM6) −3.7921, (244MMM6) −3.8432, and (334MMM6) −3.7107. The model in (37) is in fact better than the model given in Estrada's paper (statistical parameters for Estrada's structure−molar volume model with six incorrect values of edge-connectivity indices are $R = 0.993$, $S = 2.034$, and $F = 4822$).

The statistical characteristics of models based on the edge-connectivity index also support the work by Estrada and Rodriguez,[1] because one of the five physicochemical properties of octanes for which the use of the edge-connectivity index produced the best single-variable QSPR model was the molar volume. This also agrees with analyses which point out that the edge-connectivity index is more appropriate to be used in the structure−molecular volume properties modeling than the vertex-connectivity index.

The structure−molar volume models based on line graphs and edge-connectivity indices possess rather inferior statistical parameters than the models shown above:

$$\mathrm{MV} = 111.14(\pm 5.76) + 12.45(\pm 1.35)\epsilon^{[-0.50]}(\mathrm{L}) \qquad (38)$$

$$n = 69 \qquad R = 0.748 \qquad S = 11.54 \qquad F = 85$$

$$\mathrm{MV} = 67.12(\pm 3.59) + 44.49(\pm 1.67)\epsilon^{[-0.775]}(\mathrm{L}) \qquad (39)$$

$$n = 69 \qquad R = 0.956 \qquad S = 5.09 \qquad F = 714$$

Molar Refractions of Alkanes. We considered molar refractions of 69 lower alkanes also taken from Estrada.[22] Among the reported experimental values one is incorrect: Molar refraction of 34MM6 is 38.8453 instead of 43.6870.[33] We first considered the structure−molar refraction relationship using the vertex-connectivity index. The best correlation was obtained again for a rather small value of k (−0.02). The regression equation and the statistical parameters for the correlation are

$$\mathrm{MR} = 6.99(\pm 0.15) + 4.70(\pm 0.02)\chi^{[-0.02]} \qquad (40)$$

$$n = 69 \qquad R = 0.9993 \qquad S = 0.200 \qquad F = 46865$$

where MR is a short-hand notation for molar refraction. This regression equation is better than the one based on the standard value of k (−0.50):

$$\mathrm{MR} = 5.76(\pm 1.88) + 9.11(\pm 0.32)\chi^{[-0.50]} \qquad (41)$$

$$n = 69 \qquad R = 0.962 \qquad S = 1.45 \qquad F = 824$$

The model in (40) is better than, and the model in (41) is worse than, the corresponding models based on the edge-connectivity index. The best structure−molar refraction model using edge-connectivity indices was obtained for $k = -0.495$:

$$\mathrm{MR} = 7.77(\pm 0.50) + 9.26\ (\pm 0.14)\epsilon^{[-0.495]} \qquad (42)$$

$$n = 69 \qquad R = 0.992 \qquad S = 0.668 \qquad F = 4130$$

There is hardly any difference between this model and the model based on the standard value of exponent k:

$$\mathrm{MR} = 7.71(\pm 0.50) + 9.36(\pm 0.15)\epsilon^{[-0.50]} \qquad (43)$$

$$n = 69 \qquad R = 0.992 \qquad S = 0.672 \qquad F = 4090$$

Equation 43 is different from the corresponding one given by Estrada[22] as (2) in his paper. The difference is caused by erroneous values of the edge-connectivity indices for six alkanes (see the discussion above). The model in (43) is a little better than the model in the Estrada paper when the corrected values of the edge-connectivity indices are used. We also carried out the statistical analysis of Estrada's structure−molar refraction model with six incorrect values of edge-connectivity indices and obtained different statistical parameters ($R = 0.983$, $S = 0.964$, and $F = 1969$) from those reported ($R = 0.9913$, $S = 0.698$, and $F = 3782$).

The model in (43), being better than the model in (41), supports the claim by Estrada and Rodriguez[1] regarding modeling the molar refraction. In their work one of the five physicochemical properties of octanes for which the use of the edge-connectivity index produced the best single-variable QSPR model was also the molar refraction. However, when the models based on vertex- and edge-connectivity indices with variable exponents are considered, the reverse is true: the model in (40) is better than the model in (42). The model in (40) is also better than the model in (43).

Structure−molar refraction models based on alkane line graphs and edge-connectivity indices are again as in the case of structure−molar volume models inferior to models based on connectivity indices computed for alkane trees:

$$\mathrm{MR} = 11.78(\pm 1.29) + 12.30(\pm 0.56)\epsilon^{[-0.75]}(\mathrm{L}) \quad (44)$$
$$n = 69 \quad R = 0.936 \quad S = 1.861 \quad F = 474$$

$$\mathrm{MR} = 23.29(\pm 1.68) + 3.91(\pm 0.39)\epsilon^{[-0.50]}(\mathrm{L}) \quad (45)$$
$$n = 69 \quad R = 0.772 \quad S = 3.358 \quad F = 99$$

Model (40), in which the value of the exponent is rather low,[34] supports the use of the structure-molecular refraction model based on the simplest possible topological index, the number of carbon atoms V:

$$\mathrm{MR} = 2.60\ (\pm 0.18) + 4.55\ (\pm 0.02)\ \mathrm{V} \quad (46)$$
$$n = 69 \quad R = 0.999 \quad S = 0.208 \quad F = 43200$$

CONCLUSIONS

We investigated five structure−property models of alkanes. The properties considered were molecular surface-dependent properties (boiling points and gas chromatographic retention indices) and molecular volume-dependent properties (molar volumes and molar refractions). Alkanes were represented by trees and the corresponding line graphs. The vertex- and edge-connectivity indices were used as structural parameters. In each studied case we computed connectivity indices with an optimum exponent and with a standard value of −0.5. In total we generated six QSPR models for each property. The obtained results lead us to conclude the following.

(i) In all cases QSPR models based on connectivity indices with optimum exponents have better statistical parameters than the models based on connectivity indices with the standard value of the exponent (−0.5). This is fully in agreement with our earlier study[3] and the ideas of Altenburg,[35] Randić et al.,[25] and Estrada.[36] Therefore, we suggest that the modified versions of vertex- and edge-connectivity indices should be routinely employed in the structure−property modeling rather than the standard versions of the connectivity indices.

(ii) In the five cases that we studied the structure−boiling point models for octanes and nonanes and the structure−chromatographic retention index model for alkanes based on vertex-connectivity indices are better than the corresponding models based on edge-connectivity indices. Thus, it appears that the vertex-connectivity index is more appropriate to be used in the structure−molecular surface properties modeling than the edge-connectivity index. Consequently, the vertex-connectivity index may be considered as a molecular surface descriptor.

(iii) In the five cases that we studied the structure−molar volume and the structure−molar refraction models for C_5−C_9 alkanes based on the edge-connectivity index produced the best single-variable model. This agrees with the findings of Estrada and Rodriguez[1] and is suggestive that the edge-connectivity index is the better descriptor to be used in the structure−molecular volume properties modeling than the edge-connectivity index. Thus, the edge-connectivity index may be regarded as a molecular volume descriptor. The edge-connectivity index appears to be a promising molecular descriptor,[1,10,37−39] especially if the long-range contributions to this index are included in the modeling.[9,11]

(iv) The use of line graphs in this study did not improve the predictive power of the connectivity indices. Only in the case of structure−boiling point modeling for nonanes the model based on the nonane line graphs produced the best model among the possibilities considered. Since the construction of the line graphs is not difficult and the computation of their descriptors can be easily carried out, it is also reasonable to use them in the QSPR modeling, but to establish the usefulness of the line graph model in the structure−property studies, more work is needed.

ACKNOWLEDGMENT

This paper was supported, in part, by the Ministry of Science and Technology of the Republic of Croatia through Grant Nos. 00980606 (S.N. and N.T.) and 079301 (D.A. and D.B.). Part of this work was completed during the stay of S.N. and N.T. at the Natural Resources Research Institute (NRRI), University of Minnesota, Duluth, MN. S.C.B. and S.N. and N.T. (during their stay at NRRI) were supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force. This paper is contribution number 267 from the Center for Water and the Environment, NRRI. We thank the reviewers for helpful comments. We also thank Dr. Christoph Rücker (Freiburg) for helpful comments.

REFERENCES AND NOTES

(1) Estrada, E.; Rodriguez, L. Edge-Connectivity Indices in QSPR/QSAR Studies. 1. Comparison to Other Topological Indices in QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037−1041.

(2) Gutman, I.; Popović, Lj.; Estrada, E.; Bertz, S. H. The Line Graph Model. Predicting Physico-Chemical Properties of Alkanes. *ACH-Models Chem.* **1998**, *135*, 147−155 and references therein.

(3) Amić, D.; Bešlo, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. The Vertex-Connectivity Index Revisited. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 819−822 and references therein.

(4) Katritzky, A. R.; Karelson, M.; Lobanov, V. S. QSPR as a Means of Predicting and Understanding Chemical and Physical Properties in Terms of Structure. *Pure Appl. Chem.* **1997**, *69*, 245−248.

(5) Trinajstić, N. *Chemical Graph Theory*, 2nd revised ed.; CRC Press: Boca Raton, FL, 1992.

(6) Diudea, M.; Horvath, D.; Bonchev, D. Molecular Topology. 14. Molord Algorithm and Real Number Subgraph Invariants. *Croat. Chem. Acta* **1995**, *68*, 131−148.

(7) Estrada, E.; Guevara, N.; Gutman, I.; Rodriguez, L. Molecular Connectivity Indices of Iterated Line Graphs. A New Source of Descriptors for QSPR and QSAR Studies. *SAR QSAR Environ. Res.* **1998**, *9*, 229−240.

(8) Estrada, E.; Guevara, N.; Gutman, I. Extension of Edge Connectivity Index. Relationships to Line Graph Indices and QSPR Applications. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 428−431.

(9) Estrada, E. Edge-Connectivity Indices in QSPR/QSAR Studies. 2. Accounting for Long-Range Bond Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1042−1048.

(10) Estrada, E. Novel Strategies in the Search of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 279−306.

(11) Estrada, E. Connectivity Polynomial and Long-Range Contributions in the Molecular Connectivity Model. *Chem. Phys. Lett.*, in press.

(12) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vol. I, p 17.

(13) Bertz, S. H. The Bond Graph. *J. Chem. Soc., Chem. Commun.* **1981**, 818−820.

(14) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(15) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure−Activity Rela-

tionships. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73–116.

(16) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 21–57.

(17) Gutman, I.; Trinajstić, N. Graph Theory and Molecular Orbitals. III. Total π-Electron Energy of Alternant Hydrocarbons. *Chem. Phys. Lett.* **1972**, *17*, 535–538.

(18) Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399–3405.

(19) Gutman, I.; Estrada, E. Topological Indices Based on the Line Graph of the Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 541–543.

(20) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(21) Randić, M.; Jeričević, Ž.; Sabljić, A.; Trinajstić, N. On the Molecular Connectivity and π-Electronic Energy of Polycyclic Hydrocarbons. *Acta Phys. Pol.* **1988**, *74*, 317–330.

(22) Estrada, E. Edge Adjacency Relationship and a Novel Topological Index Related to Molecular Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.

(23) Trinajstić, N. *Chemical Graph Theory*, 2nd revised ed.; CRC Press: Boca Raton, FL, 1992; Chapter 4.

(24) Randić, M.; Trinajstić, N. Viewpoint 4—Comparative Structure–Property Studies: The Connectivity Basis. *J. Mol. Struct.: THEOCHEM* **1993**, *284*, 209–221.

(25) Randić, M.; Hansen, P. J.; Jurs, P. C. Search for Useful Graph Theoretical Invariants of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 60–68.

(26) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332–2339.

(27) Randić, M.; Trinajstić, N. Isomeric Variations in Alkanes: Boiling Points of Nonanes. *New J. Chem.* **1994**, *18*, 179–189.

(28) Lučić, B.; Trinajstić, N. New Developments in QSPR/QSAR Modeling Based on Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45–62.

(29) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788–802.

(30) Bošnjak, N.; Mihalić, Z.; Trinajstić, N. Application of Topographic Indices to Chromatographic Data: Calculation of the Retention Indices of Alkanes. *J. Chromatogr.* **1991**, *540*, 430–440 and references therein.

(31) Randić, M. The Structural Origin of Chromatographic Retention Data. *J. Chromatogr.* **1978**, *161*, 1–14.

(32) Canfield, E. R.; Robinson, R. W.; Rouvray, D. H. Determination of the Wiener Molecular Branching Index for the General Tree. *J. Comput. Chem.* **1985**, *6*, 598–609.

(33) E-mail of Dr. Christoph Rücker (Freiburg) to us on March 2, 2000. Dr. Rücker was informed about this by Dr. Ernesto Estrada on March 24, 1999.

(34) Note that for alkanes $\sum_{\text{edges}} [d(v_i)\ \mathrm{d}(v_j)]^0 + 1 = \mathrm{V}$.

(35) Altenburg, K. Eine Bemerkung zu dem Randićschen "Molekularen Bindungs-Index (Molecular Connectivity Index)". *Z. Phys. Chem. (Leipzig)* **1980**, *261*, 389–393.

(36) Estrada, E. Graph Theoretical Invariant of Randić Revisited. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1022–1025.

(37) Nikolić, S.; Trinajstić, N.; Baučić, I. Comparison between the Vertex- and Edge-Connectivity Indices for Benzenoid Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 42–46.

(38) Nikolić, S., Trinajstić, N.; Ivaniš, S. The Connectivity Indices of Regular Graphs. *Croat. Chem. Acta* **1999**, *72*, 875–883.

(39) Cash, G. G. Correlation of Physicochemical Properties of Alkylphenols with Their Graph-Theoretical ϵ Parameter. *Chemosphere* **1995**, *31*, 4307–4315.

CI990119V

APPENDIX 1.11 Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences

Simple Numerical Descriptor for Quantifying Effect of Toxic Substances on DNA Sequences

A. Nandy[†] and S. C. Basak*

Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway,
Duluth, Minnesota 55811

Received September 9, 1999

Many chemicals are known to be toxic to living organisms, inducing mutations and deletions at the chromosomal and genetic level. One of the tasks in risk assessment of genotoxic chemicals is to devise a simple numerical descriptor which may be used to quantify the relationship between chemical dose and the effect on the genetic sequences. We have developed numerical descriptors to characterize different DNA sequences which are especially useful in sequence comparisons. These descriptors have been developed from a graphical representational technique that enables easy visualization of changes in base distributions arising from evolutionary or other effects. In this paper we propose a scheme to use these descriptors as a label to help quantify the potential risk hazard of chemicals inducing mutations and deletions in DNA sequences.

INTRODUCTION

The deleterious effects of many chemicals and newly synthesized compounds on human and environmental health is of serious concern. Many of these chemicals are known to pass through cell barriers and cause mutations and deletions in DNAs. Recent studies have demonstrated how many common chemicals cause such effects: exposure to common environmental chemicals such as nitropyrenes present in diesel exhausts cause mutations and homologous recombinations in DNAs leading to carcinogenesis;[1,2] some polycyclic aromatic hydrocarbons from coal burning for industry and home heating form DNA adducts that have been shown to act as transplacental carcinogens and developmental toxicants[3] or induce mutations at the GC and the AT base pairs of the *hrpt* genes;[4] other chemicals such as ethylnitrosourea and ethyl methanesulfonates have been shown to induce mostly transition types of mutations in DNAs leading to chromosomal aberrations.[5] A carbonyl compound, acrolein, present in the environment as commonly used industrial chemicals, natural products, environmental contaminants and products of endogenous metabolism in human beings, has been found to cause mutations and intrastrand cross-links between guanine residues,[6] and similar effects of many other compounds are known in the literature (see, e.g., refs 7 and 8). DNA damage is also induced by excesses of heavy metals such as Rh[9] and Cu(II),[10,11] which preferentially induce depletion of guanine residues. Table 1 gives a brief list of some of the data available in recent literature on effects of chemical substances on DNA sequences.

One of the prime tasks in risk assessment of these and other chemicals and ions is to define one or more numerical descriptors of the chemical dose and the measured effect. Much of the data to date, however, consist of measures of types of mutations and deletions observed in specific genes at various levels of chemical dosages, and much of it is order of magnitude indications of genetic risk.[8] While some chemicals would induce mutations and deletions at sites with specific base pair combinations, others could lead to oxidative damages and mutations at random at intragenic and intergenic segments including point mutations and small deletions. Techniques of unbiased measures of such alterations in a DNA sequence from a set of numerical descriptors would be essential in assessing, in a universal and standard manner, the risk potential of such chemicals and form a vital link in integrating pharmacokinetics and mutational studies.

In this paper we outline such a measure arising from descriptors of DNA sequences of any specified length and show that small changes due to random point mutations or deletions in such sequences can be quantified for scaling purposes. It has developed out of a technique for graphical representation of DNA sequences but can now be done rapidly and accurately using computer programs bypassing the graphical stage altogether.

METHOD

The fundamental basis of our proposed quantitative descriptor is analysis of base distribution in a sequence by taking a running account of compositional differences in pairs of bases, e.g. intra-purines and intra-pyrimidines, as we read down the sequence from the 5′- to the 3′-end. This is most easily visualized in terms of a two-dimensional graphical representation described below. Since the method depends on small differences between the numbers of bases present in the sequences, it is very sensitive to small changes in base composition and distribution patterns.

The method of representing DNA sequences graphically using a two-dimensional Cartesian coordinate system has been explained elsewhere.[12,13] The shapes of these DNA

* To whom correspondence should be addressed. E-mail: sbasak@wyle.nrri.umn.edu.

[†] On leave from: Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta 700 032, India. E-mail: anandy43@yahoo.com.

10.1021/ci990117a CCC: $19.00 © xxxx American Chemical Society
Published on Web 00/00/0000 PAGE EST: 4.5

Table 1. Effects of Different Chemicals on DNA Sequences (Recent Studies)[a]

		mutation composition			
			substitutions (%)		
chemical	DNA sample	deletions (%)	transitions	transversions	refs and remarks
acrolein	SupF gene	24	21	55 (~GC to TA)	4
ethylnitrosourea	lacZ	5	43 (~GC to AT)	52 (~AT to TA)	5
ethylmethanosulfonate	lacZ	8	74 (~GC to AT)	18 (GC to TA)	5
heavy metals—Rh	oligomeric DNA duplexes	100 (5′-G deleted in 5′GG-3′ doublets)			9 (long-range electron transfer)
5-nitroimidazoles	Bacteroides fragilis			100 (majority C to G, CG to AT)	7
1,3-butadiene	Various—in mice, rat, humans				8 genetic hazard exists at permitted concns mutation data not available
polycyclic aromatic hydrocarbons	hprt gene	~25		~55	4

[a] Notes: The "~GC to AT" implies that the majority transitions are of the GC to AT type, etc. Acrolein is one of the a,b-unsaturated carbonyl compounds present in the environment. Nitroimidazoles, Metronidazole and dimetridazole are used in treatment of intraabdominal, pulminory, and brain abscesses and other diseases. 1,3-Butadiene is widely used in the petroleum industry.

graphs depend on the base distribution in the sequence. The plot of a typical representation is generated by moving one step in the positive x-direction for a guanine (G) in the sequence, the negative x-direction for an adenosine (A), the positive y-direction for a cytosine (C), and the negative y-direction for a thymine (T), the succession of such steps producing a graphical shape characteristic of the sequence. This essentially plots the progressive differences in the instantaneous individual totals of guanine and adenosine along the x-axis (i.e. $n_G - n_A$) and of cytosine and thymine along the y-axis (i.e., $n_C - n_T$); two other sets of axes can be similarly defined for a complete representation, but we use the one described here as the default axes system. We have shown[12] that for conserved genes such plots are shape similar thereby making identification of a new sequence of the gene family possible rapidly and easily by visual inspection alone; elsewhere we have shown that one can read off base preferences and local abundances directly from the shape of these graphs[14] or identify coding and noncoding regions of the sequences.[15] Changes in base distribution and composition induce changes in the visual plots of the DNA sequences; for the same genes for different species we have noticed systematic drifts in the sequence pattern which have been attributed to evolutionary changes.[16]

Differences in the plots of a family of genes can be quantitatively assessed.[17] This method consists essentially of defining a set of moments of the graph points around the origin of the plot. In the first order we define quantities $\mu_1(x)$ and $\mu_1(y)$ which are the sum of the x- and y-coordinate values of each point averaged by the total number of points in the distribution. One can then define a graph radius for each plot

$$g_R = [(\mu_1(x))^2 + (\mu_1(y))^2]^{1/2}$$

and correspondingly a distance measure between two graphs:

$$d(s,s') = [(\mu_1(x) - \mu_1(x'))^2 + (\mu_1(y) - \mu_1(y'))^2]^{1/2}$$

where s and s' represent the two graphs. We have observed[17] that small differences in DNA sequences arising out of base mutations and deletions manifest themselves in observable changes in g_R and d. We propose to use the g_R as one numerical descriptor of a sequence and deviation from g_R, Δg_R, as a measure of the changes in a sequence as a consequence of genotoxic effects of chemicals. For greater precision, one could also use a set of $\mu_1(x)$, $\mu_1(y)$, and g_R as numerical descriptors of a DNA sequence.

RESULTS AND DISCUSSIONS

As a preliminary exploration of this technique, we have used the complete human β globin gene sequence (from the HSHBB sequence of the EMBL DNA database rel 31), inclusive of the introns and exons, as the control sequence. This has a total of 1424 bases consisting of 444 (31.2%) bases in the coding regions and 980 (68.8%) in the noncoding part. Plot 1 in Figure 1a shows the graphical representation of this gene starting from exon 1 through introns 1 and 2 to exon 3. Intron 1 is G-rich and shows a horizontal shift to the right; intron 2 has a T-rich part in the initial stages, represented on our graph as an almost vertical drop, and then a long stretch of TA repeats that move the graph generally in a southwesterly direction ending with exon 3 represented as a small region of a dense cluster of points. Exons 1 and 2 are also represented as (less dense) clusters of points unlike the long runs of the introns; we have elsewhere[15] exploited this characteristic difference between intron and exon representations as a means for determining protein coding regions in new sequences.

With regard to the problem at hand, we simulated the effects of Rh and Cu(II) toxicity on a DNA by performing programmatically random deletions of several guanines in the sample sequence. Such deletions will tend to alter the $\mu_1(x)$ in the default representation with a bias toward negative x-values (because of a higher percentage of adenosines in the altered sequence) while leaving the $\mu_1(y)$ unchanged and will consequently alter the graph radius. Graphically, the reductions in the number of guanines will make the plot shift to the left in the default reference frame, and the shift will be greater for a greater degree of deletions effected. This is evident visually from a low value of 5% deletions in the complete sequence (Figure 1a). The values of Δg_R for different numbers of guanine deletions are plotted in Figure 1b.

In the case of mutations, the graph radius is quite sensitive to small changes and to specific base positions affected. A

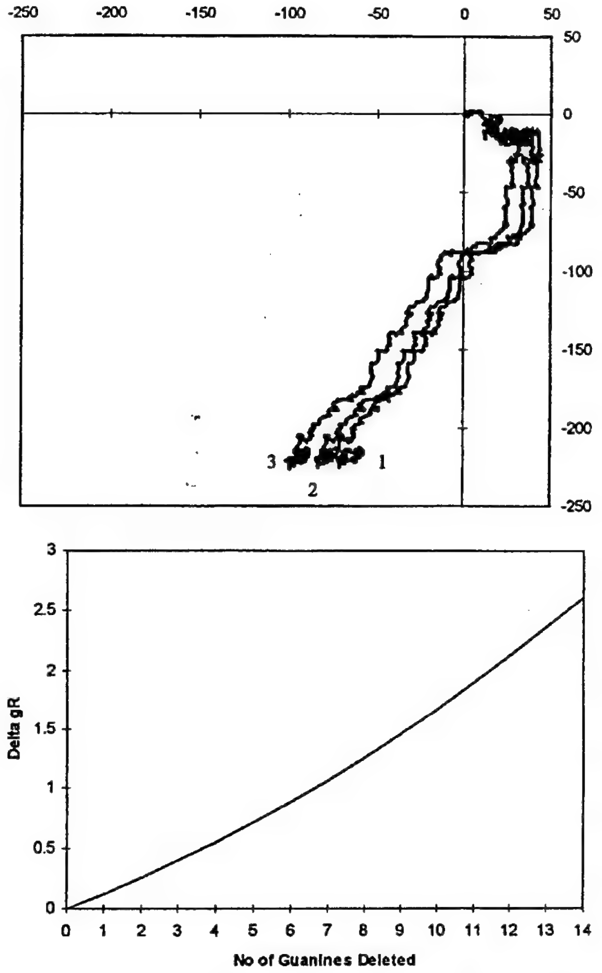


Figure 1. (a, top) Human β globin gene and its model modifications plotted in the two-dimensional representation system. Axes as explained in the text. Plot 1 is for the normal human β globin gene complete with exons and introns. Plot 2 is for the same gene with 5% random depletion in guanine residues. Plot 3 is the same gene with 10% depletion in guanine bases. (b, bottom) Plot of changes in graph radius (Δg_R) against guanine number for deletion of guanines in positions 1−14.

mutation in the first position, reading from the 5′-end, effects the maximum change while a mutation in the last base has the least effect; this is easily understood from the fact that the change in the first position alters the coordinate value of each subsequent point all the way to the last base and thus affects the value of μ_1 much more than would be the case for mutation of the last base. (The argument remains true when read from the 3′-end and as long as one is consistent in one's convention; here we use the common convention of reading from the 5′-end.) Figure 2 shows Δg_R plotted against the guanine number for mutation of one guanine to cytosine in each position of the guanine in the complete sequence of the human β globin gene. It is interesting to note that Δg_R has a unique value for each position, and, as can be expected, the value goes down to almost zero for the last guanine (the kink seen in the curve occurs at a large gap between successive guanines). Mutations of guanine to adenosine will produce smaller amount of changes in Δg_R since this is a change occurring exclusively in the x-direction and lead to a contraction or expansion of the general curve, whereas the previous mutations produced a change in direction of the plot in our default axes system. Figure 3 shows the variation of Δg_R with guanine number for mutation of a single guanine to adenosine. We have noted elsewhere[18] that Δg_R can therefore be used as quantitative descriptors for indexing single nucleotide polymorphic genes.

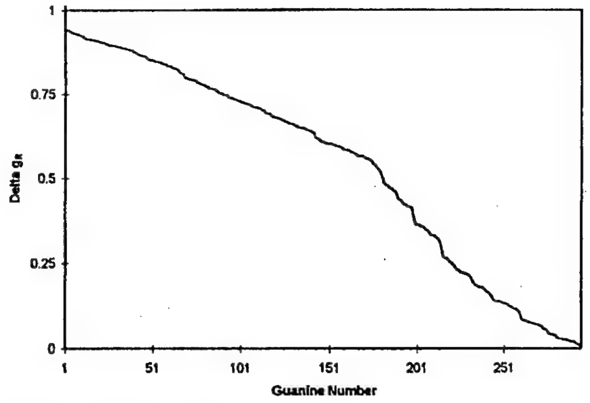


Figure 2. Plot of changes in graph radius (Δg_R) against the guanine number for mutation (G to C) of single guanine to cytosine at various positions.

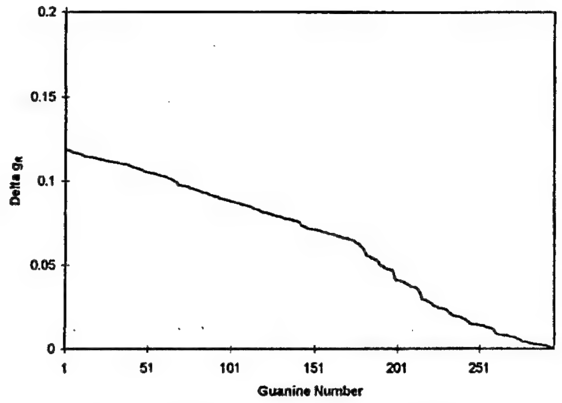


Figure 3. Plot of changes in graph radius (Δg_R) against the guanine number for mutation (G to A) of single guanine to adenosine at various positions.

In the present case of indexing as a measure of risk assessment for toxicity, the sensitivity of Δg_R raises the question of adequate knowledge of the exact location of the toxic damage. Since any random mutation or deletion could arise from the genotoxic effects, it would be preferable to average over the entire range of values of Δg_R over the chosen DNA segment to arrive at an acceptable index value for purposes of comparative assessment. For example, for the case of mutation of one guanine to adenosine, the average value of Δg_R is 0.064 while that for the case of guanine to cytosine is 0.537, and an index for the two types of causative chemicals that produce just this level of mutation could be written in thousandths as 64 or 537.

In the case of multiple base mutations also this trend of different values of Δg_R for mutations at different base positions will hold true: e.g., mutations of three guanines to cytosines will cause maximum deviation from g_R when the mutations occur in the first three guanines (Δg_R = 2.789 76 compared to the unmutated gene), and the change will be least when the mutations take place in the last three guanines (Δg_R = 0.031 41 compared to the unmutated gene). Multiple mutations will therefore create a field of values for

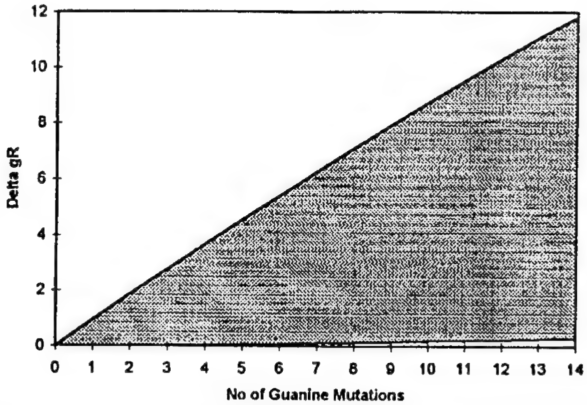


Figure 4. Plot of changes in graph radius (Δg_R) against the number of guanines mutated for G to C mutations. The upper line is the highest value and the bottom line the lowest value of Δg_R for a given number of mutations.

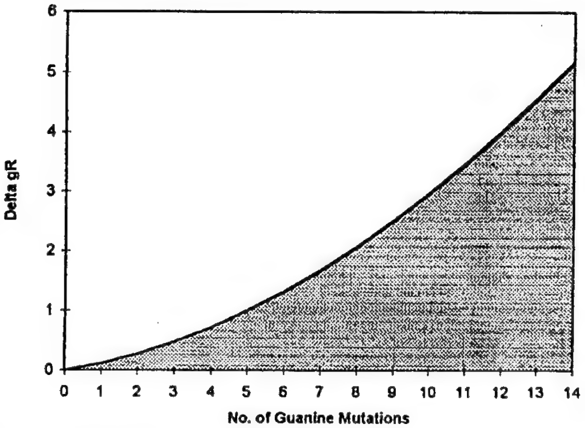


Figure 5. Plot of changes in graph radius (Δg_R) against the number of guanines mutated for G to A mutations. The upper line is the highest value and the bottom line (not visible on this graph range) the lowest value of Δg_R for a given number of mutations.

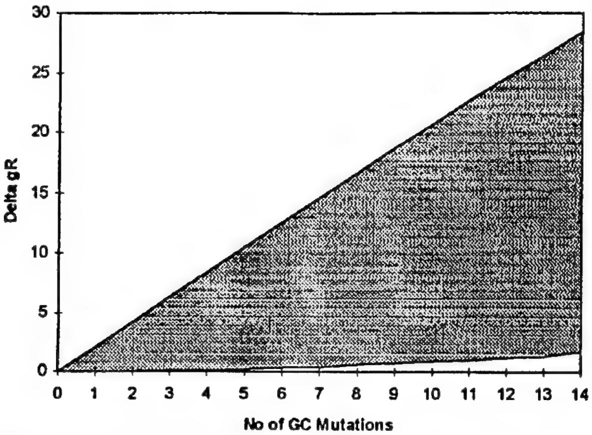


Figure 6. Plot of changes in graph radius (Δg_R) against the number of GC to AT mutations. The upper line is the highest value and the bottom line the lowest value of Δg_R for a given number of mutations.

Δg_R, the maximum for a specific number of mutations being the value realized from mutations in the first of those bases. These maximum values will thus form an envelope as shown in Figure 4, and a lower bound will be created by the minimum values of Δg_R; all values between these two boundaries will relate to the different bases in the sequence that can be mutated for any specified number of mutations. Figure 5 shows similar data for the various degrees of G to A possible mutations.

While we have discussed these effects on the hypothesis of G to C and G to A mutations, these results can be generalized to mutations in any base combinations also. For example, in the case of genetic mutations induced by high levels of toxic chemicals where more than one base can be affected, e.g. mutations of the type GC to AT shown in Figure 6, which occurs in the case of the ethylnitrosourea and ethyl methanosulfonate types of compounds, one can determine the value of Δg_R from a sample sequence exposed to a standard dosage and use that value as an index for measuring the least number of mutations that can be generated from such a number. From Figure 6, for example, it can be seen that a Δg_R of 10 implies that the number of corresponding mutations will be five GC doublets or more.

Thus an experimental measure of Δg_R for a given dose of a toxic chemical can lead to association of an index value that will permit easy gradation of chemicals on levels of toxicity. Each toxin will affect a DNA in its own unique ways: some by deleting a preferred base, some by causing random mutations in one or more preferred bases. The usefulness of an index such as Δg_R arises from associating one number with each dosage level of each chemical providing an easy path to associating risk with dosage without having to enumerate which base and how many are mutated or deleted. Δg_R thus enables a normalization approach to risk assessment of genotoxic chemicals where no other such measure is readily available.

Note that the method is not dependent on the type of DNA sequence used; while for some chemicals specific DNA segments will be susceptible to damage, for others damages can occur in any of the coding or noncoding segments as for example in case of Cu(II) and Rh induced damages. The indexing can be done for all these cases with respect to any standard sequence segment chosen.

CONCLUSION

Thus we see that the concept of graph radius in a graphical representation of a DNA sequence can be extended to make quantitative estimation of any changes in the sequence. This observation indicates that it is possible to consider using such quantitation as an index of the intensity of the effects in the case of changes arising out of effects of genotoxic chemicals. As of now, however, we are restricted by the paucity of experimental data to only indicating the use of Δg_R as a possible index; experimental work so far are generally in the nature of inquiries into the kinds of changes induced in DNA sequences by genotoxic chemicals, whereas building up a quantitative index would require controlled experiments relating dosage and the extent of DNA damage.

Our work has shown that Δg_R, the change in g_R, is a very sensitive indicator of changes in a sequence arising out of base depletions and mutations. This provides us therefore a numerical descriptor of the alterations in base distribution and composition of DNA sequences and can be used to compare with any standard or control sequence. Δg_R, therefore, averaged over its relevant range of values, can be used as a numerical descriptor to provide a measure of the genotoxic effects of chemicals such as oxidants such as Rh and Cu(II), or acrolein, ethyl methanosulfonate, or any other chemicals whose effect on DNA sequences can occur in a random manner and therefore can affect any part of the DNA whether coding or noncoding. In the case of genotoxins that affect specific genes or base combinations, the Δg_R will need to be calculated for those specific genes only, and there the sensitivity of the measure can be exploited to provide an indicator of the genotoxicity level of the chemicals.

ACKNOWLEDGMENT

This paper is contribution number 268 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grant F49620-96-1-0330 from the United States Airforce. We acknowledge gratefully the help received from Ms. Denise Mills for literature search and Mr. Brian Gute for the computer setups. A.N. would also like to thank all members of the NRRI for the kind hospitality and courtesy shown during his stay at the Institute there.

REFERENCES AND NOTES

(1) Beland, F. A. How do chemicals in diesel engine exhaust damage DNA? Health Effects Institute Research Report No. 46; Health Effects Institute: Cambridge, MA, 1991.

(2) Maher V.; Bhattacharya, N. P.; Chia-Miao, Mah. M.; Boldt, J.; Yang, J.–L.; McCormick, J. J. Relationship of nitropyrine-derived DNA adducts to carcinogenesis; Health Effects Institute Research Report No 55; Health Effects Institute: Cambridge, MA, 1993.

(3) Whyatt, R. M.; Santella, R. M.; Jedrichowsky, W.; Garte, S. J.; Bell, D. A.; Ottman, R.; Gladek-Yarborough, A.; Cosma, G.; Young, T.–L.; Cooper, T. B.; Randall, M. C.; Manchester, D. K.; Perera, F. P. Relationship between ambient air pollution and DNA damage in Polish mothers and new-borns. *Environ. Health Perspect.* **1998**, *106* (Suppl. 3), 821–826.

(4) Zanesi, N.; Mognato, M.; Pizzato, M.; Viezzer, C.; Ferri, G.; Celotti, L. Determination of hprt mutant frequency and molecular analysis of T-lymphocyte mutants derived from coke-oven workers. *Mutat. Res.* **1998**, *412*, 177–186.

(5) Suzuki, T.; Hayashi, M.; Wang, X.; Yamamoto, K.; Ono, T.; Myhr, B. C.; Sofuni, T. A comparison of the genotoxicity of ethylnitrosourea and ethyl methanesulfonate in lacZ transgenic mice (Muta Mouse), *Mutat. Res.* **1997**, *395*, 75–82.

(6) Kawanishi, M.; Matsuda, T.; Nakayama, A.; Takebe, H.; Matsui, S.; Yagi, T. Molecular analysis of mutations induced by acrolein in human fibroblast cells using supF shuttle vector plasmids, *Mutat. Res.* **1998**, *417*, 65–73.

(7) Trinh, S.; Reysset, G. Mutagenic action of 5-nitroimidazoles: In vivo induction of GC → CG transversion in two Bacteroides fragilis reporter genes. *Mutat. Res.* **1998**, *398*, 55–65.

(8) Pacchierotti, F.; Adler, I.–D.; Anderson, D.; Brinkworth, M.; Demopoulos, N. A.; Laehdetie, J.; Osterman-Golkar, S.; Peltonen, K.; Russo, A.; Tates, A.; Waters, R. Genetic effects of 1,3-butadiene and associated risk for heritable damage. *Mutat. Res.* **1998**, *397*, 93–115.

(9) Hall, D. B.; Holmlin, R. E.; Barton, J. K. Oxidative DNA damage through long-range electron transfer. *Nature* **1996**, *382*, 731–735.

(10) Richard, H.; Daune, M.; Schreiber, J. P. *Biopolymers* **1973**, *12*, 1.

(11) Foerster, W.; Bauer, E.; Schitz, H.; Akimenko, N. M.; Minchenkova, L. E.; Evolokimov, Y. M.; Varshakovsky, Y. M. *Biopolymers* **1979**, *18*, 625.

(12) Nandy, A.; A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes. *Curr. Sci.* **1994**, *66* (4), 309–314.

(13) Ray, A.; Raychaudhury, C.; Nandy, A. Novel Techniques of Graphical Representation and Analysis of DNA Sequences–A Review. *J. Biosc.* **1998**, *23* (1), 55–71.

(14) Nandy, A.; Nandy, P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Current Sci.* **1995**, *68* (1), 75–85.

(15) Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.* **1996**, *12* (1), 55–62.

(16) Nandy, A. Graphical Analysis of DNA Sequence Structure: III. Indications of Evolutionary Distinctions and Characteristics of Introns and Exons. *Current Sci.* **1996**, *70* (7), 661–668.

(17) Raychaudhury, C.; Nandy, A. Indexing Scheme and Similarity Measures for Macromolecular Sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.

(18) Nandy, A.; Nandy, P.; Basak, S. C. Quantitative Descriptor for SNP Related Gene Sequences. Manuscript in preparation.

CI990117A

APPENDIX 1.12 Topological indices: Their nature and mutual relatedness

Topological Indices: Their Nature and Mutual Relatedness

Subhash C. Basak,*[,†] Alexandru T. Balaban,[‡] Gregory D. Grunwald,[†] and Brian D. Gute[†]

Natural Resources Research Institute, University of Minnesota—Duluth, Duluth, Minnesota 55811, and
Organic Chemistry Department, Polytechnic University Bucharest, Splaiul Independentei 313,
77206 Bucharest, Romania

Received September 9, 1999

We calculated 202 molecular descriptors (topological indices, TIs) for two chemical databases (a set of 139 hydrocarbons and another set of 1037 diverse chemicals). Variable cluster analysis of these TIs grouped these structures into 14 clusters for the first set and 18 clusters for the second set. Correspondences between the same TIs in the two sets reveal how and why the various classes of TIs are mutually related and provide insight into what aspects of chemical structure they are expressing.

INTRODUCTION

A major part of the current research in mathematical chemistry, chemical graph theory, and quantitative structure–activity/property relationship studies involves topological indices. Topological indices (TIs) are numerical graph invariants that quantitatively characterize molecular structure. A graph $G = (V, E)$ is an ordered pair of two sets V and E, the former representing a nonempty set and the latter representing unordered pairs of elements of the set V. When V represents the atoms of a molecule and elements of E symbolize covalent bonds between pairs of atoms, then G becomes a *molecular graph* (or *constitutional graph*, because there is no stereochemical information). Such a graph depicts the topology of the chemical species. A graph is characterized using graph invariants. An invariant may be a polynomial, a sequence of numbers, or a single number. A numerical graph invariant (i.e., a single number) that characterizes the molecular structure is called a topological index.

OVERVIEW OF TOPOLOGICAL INDICES USED IN THE PRESENT STUDY

A large number of topological indices have been defined and used.[1–11] The majority of TIs are derived from the various matrices corresponding to molecular graphs. The adjacency matrix $\mathbf{A}(G)$ and the distance matrix $\mathbf{D}(G)$ of the molecular graph G have been most widely used in the formulation of TIs. Integer-number local vertex invariants (LOVIs) are the vertex degrees (v_i) and the distance sums (distasums, d_i) resulting from summation over rows or columns of entries in the adjacency and distance matrices, respectively. By mathematical operations performed on such LOVIs, one can obtain a molecular descriptor, *i.e.*, a topological index. Wiener's index W (eq 1),[2] the Zagreb group index M_1 (eq 2),[11] Randić's connectivity index, χ (eq 3),[4] the higher-order connectivity indices, ${}^n\chi$, for paths of length n defined by Kier and Hall,[5] and the J index (eq 4)[6] fall in this category.

$$W = (\textstyle\sum_i d_i)/2 \tag{1}$$

$$M_1 = \textstyle\sum_i v_i^2 \tag{2}$$

$$\chi = \textstyle\sum_{ij}(v_i v_j)^{-1/2} \tag{3}$$

$$J = [q/(\mu + 1)]\textstyle\sum_{ij}(d_i\, d_j)^{-1/2} \tag{4}$$

The summations in formulas 3 and 4 are over all edges $i-j$ in the hydrogen-depleted graph. The numbers q of graph edges and μ of cycles in the graph are introduced into formula 4 in order to avoid the automatic increase of J with graph size and cyclicity. Indeed, for an infinite linear carbon chain it was demonstrated that $J = \pi = 3.14159$. The nature of atoms can be taken into account by means of parameters based on their relative atomic numbers, electronegativities, or covalent radii, with respect to those of carbon atoms, multiplying the corresponding distasum in formula 4 for J.

The mean-square-root distance D derived from all topological distances (denoted by i in the next formula) is defined as[6b]

$$D = [(\textstyle\sum_i i^2)/(\sum_i i)]^{1/2} \tag{5}$$

For taking into account the chemical nature of atoms symbolized by vertices, Kier and Hall advocated the use of "valence connectivity indices".[5a,b] These are calculated with formulas similar to Randić's (eq 3), but products of edge end point (or path vertex) invariants are no longer of vertex degrees but of weights (valence delta values δ_i) given by formula 5

$$\delta_i = (Z_i^{\mathrm{v}} - H_i)/(Z_i - Z_i^{\mathrm{v}} - 1) \tag{6}$$

where Z_i^{v} stands for the number of valence electrons in atom i, Z_i is its atomic number, and H_i is the number of hydrogen atoms attached to atom i.

The most recent additions to the Kier–Hall armamentary of TIs are electrotopological state indices.[5c]

* Corresponding author. Tel: (218)720-4230. Fax: (218)720-4328. E-mail: sbasak@nrri.umn.edu.

[†] University of Minnesota.

[‡] Polytechnic University Bucharest.

10.1021/ci990114y CCC: $19.00 © 2000 American Chemical Society
Published on Web 02/18/2000

Another class of molecular descriptors, the information-theoretic indices, are derived from an entirely different reasoning. In this case, the complexity or mode of partitioning of structural features is decomposed into disjoint subsets using an equivalence relation; a molecular complexity index is then computed using Shannon's idea of information content or complexity.[12] Real-number local vertex invariants (LOVIs), on the other hand, may also be defined starting from different matrices other than **A**(G) or **D**(G) or by applying information theory at the vertex level. Thus, topological indices U, V, X, and Y were defined.[13] Bonchev and Trinajstić described several information-theoretic TIs reviewed thoroughly in Bonchev's book.[7]

The information-theoretic indices developed by Basak and co-workers take into account all atoms in the constitutional formula (hydrogens also being included), and one considers the information content provided by various classes of atoms based on their topological neighborhood. There are three main types of informational indices developed by Basak et al.: IC (mean information content or complexity of a hydrogen-filled graph, with vertices grouped into equivalence classes having r vertices; the equivalence is based on the nature of atoms and bonds, in successive neighborhood groups); CIC (complementary information content); and SIC (structural information content), and they are not inter-correlated with other TIs. In the following formula, the summation spans the range from $i = 1$ to $i = r$:

$$\mathrm{IC}_r = -\sum_i p_i \log_2 p_i \tag{10}$$

$$\mathrm{SIC}_r = \mathrm{IC}_r/\log_2 N \tag{11}$$

$$\mathrm{CIC}_r = \log_2 N - \mathrm{IC}_r \tag{12}$$

The probability that a randomly selected vertex occurs in the ith equivalence class is denoted by p_i. The IC_r, SIC_r, and CIC_r indices can be calculated for different orders of neighborhoods, r ($r = 0, 1, 2, ..., \rho$), where ρ is the radius of the molecular graph G. At the 0th-order level, the atom set is partitioned solely on the basis of its chemical nature; at the level of the first-order topological neighborhood, the atoms are partitioned into disjoint subsets on the basis of their chemical nature and their first-order bonding topology. At the next level, the atom set is decomposed into equivalence classes using their chemical nature and bonding pattern up to the second-order bonded neighbors. The process is continued until consideration of higher-order neighbors does not yield further increase in the number or composition of disjoint subsets.

A large variety of real-number local vertex invariants, and thence a larger variety of TIs, were described on the basis of converting a matrix (**A** or **D** for instance) into a system of linear equations. This is done by means of two column vectors that can convey topological, chemical, or numerical information. One nonzero vector is the free term of the system of equations. The other one (which may be zero, but this restricts the choices on available supplementary information) becomes the main diagonal of the matrix (if both vectors were zero, then some negative LOVIs would result with difficulties of interpretation). These vectors may be the following integers: Z (atomic number of the atom corresponding to each vertex), V (vertex degree), I (identity), N (number of non-hydrogen atoms, or order of the graph), N^k (power k of N). Less frequently, one may use for periodicity of chemical properties real numbers: S (electronegativity) or R (covalent radius) of the atom corresponding to each vertex. The resulting matrix with the vector for the main diagonal constitutes the set of coefficients for the N unknowns that represent the real-number LOVIs of the N vertices. The triplet (matrix, vector for the main diagonal and vector for the free term) also serves as notation for LOVIs and for the derived TIs. After the system of N linear equations is solved, the LOVIs (x_i) are assembled into a "triplet TI" based on one of the following operations:

1. summation, $\Sigma_i x_i$;
2. summation of squares, $\Sigma_i {x_i}^2$;
3. summation of square roots, $\Sigma_i {x_i}^{1/2}$;
4. sum of inverse square root of cross-product over edges ij, $\Sigma_{ij} x_i x_j)^{-1/2}$;
5. product, $N[\Pi_i x_i]^{1/N}$.

Numbers 1–5 of the above operations after the triplet complete the notation of the triplet TIs.[14]

To conclude this brief review of TIs, one should mention recent progress that includes other matrices such as the reciprocal distance matrix that yields Harary indices,[15] the regressive distance matrices,[16] the Szeged matrix,[17] and the resistance distance matrix that affords Kirchhoff indices.[18] So-called optimal structural descriptors can be obtained from some TIs by varying some parameters and thereby adapting them to the database;[19] alternatively, in Randić-type formulas (eqs 3, 4) the exponent is allowed[20] to differ from $^1/_2$. Three-dimensional molecular descriptors can be derived from geometrical and topological structural features of molecules.[21]

Each of the indices above-discussed is a "global" parameter; i.e., it quantifies certain aspects of the entire molecular structure using a single number.

It is clear from the above discussion that the set of TIs is a group of heterogeneous entities. They have been defined to characterize molecular structure on the basis of distinct objectives and motivations. Despite their distinctive characteristics, TIs share certain common features. A topological index maps a set of chemicals C into the set R of real or integer numbers. Therefore, TIs quantify some general aspects of molecular architecture such as size, shape, symmetry, bonding type, cyclicity, branching pattern, etc.

Topological indices have been used for isomer discrimination, quantification of the structural similarity/dissimilarity of molecules, and prediction of property/activity from structure.[19] The widespread use of TIs obviously encourages one to ask some fundamental questions about them: What is the fundamental nature of TIs? To what degree are they intercorrelated? How does one extract orthogonal information from TIs?

The intercorrelation of TIs was studied earlier with a limited set of invariants. Thus, Motoc and Balaban[22] described graphically the intercorrelations of the few TIs known until 1981. These aspects were reviewed in the early 1980s.[23] Basak et al. studied the mutual relatedness of a set of 90 TIs calculated for a set of 3692 diverse chemicals.[24] A third study by Todeschini et al. will be discussed in the last section of this paper.

All such studies were limited in the sense that they analyzed data on a smaller and less diverse group of TIs. Therefore, in this paper, we have studied the mutual

Table 1. Summary of Chemical Classes or Features in Databases Analyzed

chemical classes or features	database A (hydrocarbons)	database B (diverse)
total number of compounds	139	1037
hydrocarbons	139	565
alkanes, cyclic alkanes	73	206
aromatics	66	288
alkyl benzenes	29	80
fused rings	37	56
polycyclic aromatics	37	49
non-hydrocarbons	0	472
halogen-containing compounds		359
heteroatom-containing compounds (sulfur or phosphorus)		101
Compounds containing both halogens and heteroatoms		12
organosulfides		105
organophosphorus		8

relatedness of a set of 202 TIs. We have also tried to extract useful and orthogonal structural information from the calculated TIs. This study also reports, for the first time, a comprehensive discussion of Basak's information content indices (IC_r, SIC_r, CIC_r), the triplet indices (proposed by one of the present authors), and Balaban's average distance-based connectivity index J as compared to the traditional and more widely used indices.

The goal of this paper is two-fold: (a) to study the degree of intercorrelation among the various types of topological indices and (b) to extract mutually uncorrelated (orthogonal) topological parameters that can be used for QSAR/QSPR studies, quantitation of intermolecular similarity/dissimilarity, and characterization of real and virtual combinatorial libraries. To this end, we studied the mutual relatedness of a set of more than 200 topological indices in this paper.

METHODS

Chemical Databases. There were two sets of chemicals analyzed in this study: a set of 139 hydrocarbons to represent a moderately homogeneous set of chemicals and a set of 1037 diverse chemicals. The hydrocarbons consisted of 73 C3−C9 alkanes, 29 alkylbenzenes, and 37 polycyclic aromatic hydrocarbons.[25] The diverse set of 1037 compounds consists of those chemicals from the U.S. EPA ASTER system[26] for which a measured boiling point was available and hydrogen-bonding potential (as measured by HB1 = 0) did not exist. The composition of these data sets is indicated in Table 1. Table 2 presents the list of all 202 parameters calculated in this study.

Calculation of TIs. The TIs calculated for this study (some of which are included in Table 2) include Wiener number W,[2] molecular connectivity indices as calculated by Randić[4] and Kier and Hall,[5] frequency of path lengths of varying size,[5] information-theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[7] Roy et al.,[27] Basak et al.,[28−31] and Raychaudhury et al.,[32] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[28−32]

Table 2. Symbols and Definitions of Topological Parameters

index	definition
$\underline{I}^{W}{}_{D}$	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}^{W}{}_{D}$	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^{D}	degree complexity
H^{V}	graph vertex complexity
$\underline{H^{D}}$	graph distance complexity
$\overline{\mathrm{IC}}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
M_1	a Zagreb group parameter = sum of square of degree over all vertices
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi$	path connectivity index of order $h = 0-6$
${}^{h}\chi$	cluster connectivity index of order $h = 3-6$
${}^{h}\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
${}^{h}\chi_{Ch}$	chain connectivity index of order $h = 3-6$
${}^{h}\chi^{b}$	bond path connectivity index of order $h = 0-6$
${}^{h}\chi^{b}{}_{C}$	bond cluster connectivity index of order $h = 3-6$
${}^{h}\chi^{b}{}_{Ch}$	bond chain connectivity index of order $h = 3-6$
${}^{h}\chi^{b}{}_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
${}^{h}\chi^{v}$	valence path connectivity index of order $h = 0-6$
${}^{h}\chi^{v}{}_{C}$	valence cluster connectivity index of order $h = 3-6$
${}^{h}\chi^{v}{}_{Ch}$	valence chain connectivity index of order $h = 3-6$
${}^{h}\chi^{v}{}_{PC}$	valence path-cluster connectivity index of order $h = 4-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^{B}	Balaban's J index based on bond types
J^{X}	Balaban's J index based on relative electronegativities
J^{Y}	Balaban's J index based on relative covalent radii
triplet	Global invariants based on solutions of linear equation systems using the adjacency matrix (**A**), distance matrix (**D**), and column/row vectors: distance sums (S), atomic number (Z), number of non-hydrogen atoms (N and N^2), vertex degree (V), or numerical constants (1). Notation is described by triplets (e.g., AZV). Results are weightings for each atom in a molecule. These weights are combined by five possible formulas: 1 = sum of weights, $\sum_i x_i$; 2 = sum of squared weights $\sum_i x_i^2$; 3 = sum of square root of weights $\sum_i x_i^{1/2}$; 4 = sum of cross-products $\sum_i (x_i \cdot x_j)^{-1/2}$; and 5 = product of weights $N \cdot [\sum_i x_i]^{1/N}$.

and Balaban's *J* indices[6] as well as triplet indices.[14] The majority of the TIs were calculated using the program POLLY 2.3.[33] The *J* indices and triplet indices were calculated using software developed in-house by the authors.

STATISTICAL ANALYSIS

For both sets of chemicals, the computed TIs were transformed by the natural logarithm of the index plus a constant, generally 1. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices.

For each set, a technique known as variable clustering was performed using the SAS procedure VARCLUS.[34] The variable-clustering procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. This is accomplished by a repeated principal-components analysis of the sets of indices. The initial principal-component analysis examines all indices and defines two principal components or eigenvectors. If the eigenvalue for the second component is >1.0, the indices are split into separate clusters by correlating the indices with the first and second principal components. Those indices most correlated with the first component form one cluster and those indices most correlated with the second component form another cluster, thus forming two disjoint clusters. A principal-component analysis is then performed for each cluster of indices, with the cluster being split if the eigenvalue for the second component is >1.0. The procedure is repeated until the second eigenvalue is <1.0 for all clusters.

RESULTS AND DISCUSSION

The first database (denoted by A) consists of 139 hydrocarbons (alkanes, alkylbenzenes, and polycyclic aromatics) and 162 TIs. The number of indices examined was reduced from the original 202 by removing all but one of the degenerate (i.e., correlation of 1.0) indices and those indices that were constant (0.0) for all chemicals. The second database (denoted by B) is a diverse one and contains 1037 chemical structures and 176 nondegenerate, nonconstant indices.

The results of the variable-cluster analysis will be presented, first discussing how the descriptors (variables) for database A become clustered and then surveying the descriptor clustering for database B, as well as the correspondence between these clusters. Intercluster correlation will then be described.

The clusters have been ordered according to decreasing numbers of descriptors in each cluster; when clusters contain the same number of descriptors, the numbering of the corresponding clusters is arbitrary.

In Figure 1, one can see, in graphical form, on the left-hand side the points denoting the clusters that group together the descriptors for the hydrocarbon database A and on the right-hand side those corresponding to the diverse database B. Each cluster is denoted by a letter (A or B) and a number. The total number of variables in each cluster is written under each point. Full lines connect A-type with B-type clusters, having inscribed on them the numbers of descriptors common to each pair of clusters; when no number is inscribed, this indicates a single common descriptor. Dashed side lines denote the descriptors that do not have counterparts in the other set of clusters, and the associated numbers on these side lines indicate the numbers of such "orphan" descriptors. Because the two data sets differ both in the numbers of compounds and in their structures, it is normal to expect that clusters for one data set will have counterparts in several clusters in the other data set. This is indeed what was found to happen, as will be shown below when the diverse data set will be analyzed.

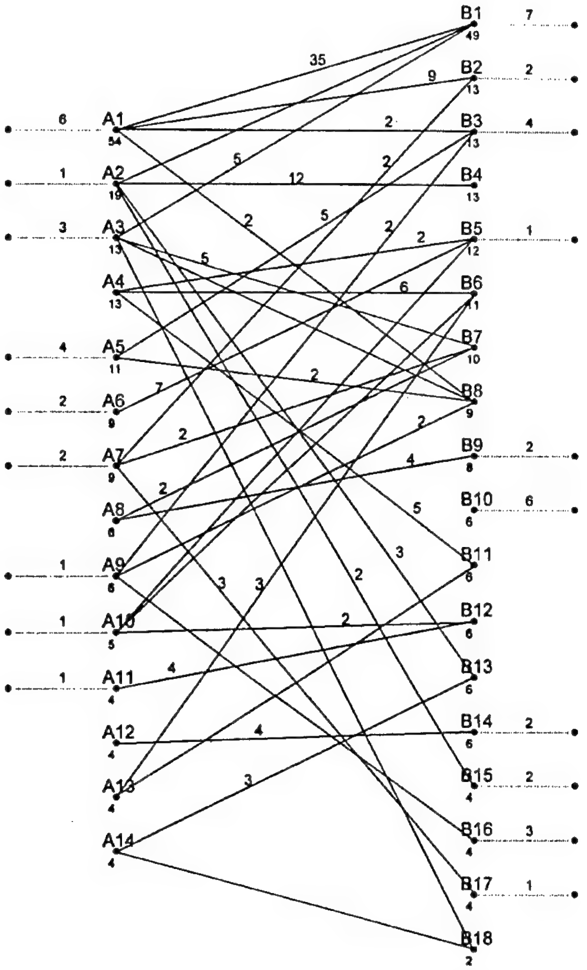


Figure 1. Associations between clusters of descriptors for the hydrocarbon database (A-type clusters) and the database with diverse compounds (B-type clusters). Solid lines connect A-type descriptors with B-type descriptors, and the numbers of common descriptors are indicated on such lines (when no number is indicated, there is just one common descriptor). Dashed lateral lines indicate descriptors that have no correspondence for the other type.

Only in a single case have we found a one-to-one correspondence between clusters of descriptors corresponding to the two data sets (A12 and B14). Nevertheless, in several instances (A6, A11; B4, B9, B15, B16, and B17), a cluster for one data set (say, A) was found to have all its descriptors in common with only one cluster of the other data set (say, B); however, this latter cluster also contains descriptors found in more than one cluster of the other set.

Clustering of Descriptors for Hydrocarbons. The descriptors for database A are grouped in 14 clusters summarized in Table 3. Cluster A1 has 54 from the total of 162 descriptors; therefore, it groups together about one-third of all variables. These descriptors depend on both the shape and the size (magnitude) of the molecular graph; such

Table 3. Summary of Variable Clustering for 139 Hydrocarbons

cluster	number of variables	representative variables (max. 25% of total listed)
A1	54	DN^2Z_4, DN^2N_4, P_0, AZV_4, ASZ_4, ANN_3, ANN_5, AZN_3
A2	19	$^6\chi$, P_7, $^5\chi$, $^6\chi^b$, $^6\chi^v$
A3	13	$^0\chi^b$, $^0\chi^v$, ANZ1
A4	13	SIC_6, SIC_5, IC_6
A5	12	DSZ_1, DSZ_5, ASZ_1
A6	9	DSZ_3, DSN_5
A7	9	DSN_3, DN^2N_1
A8	6	$^5\chi^v{}_C$, $^5\chi^b{}_C$
A9	6	DSZ_2, ASZ_2
A10	5	SIC_1
A11	4	CIC_1
A12	4	$^3\chi^v{}_C$
A13	4	SIC_3
A14	4	$^5\chi_{Ch}$

descriptors include the Randić connectivity index, the Kier–Hall simple path connectivity indices, the Zagreb group indices, and many triplet indices having as the main diagonal column vector the atomic number Z or the total number N of vertices.

Cluster A2 with about $^1/_8$ of the total number of descriptors includes molecular connectivity indices of order higher than 5, the J indices, and two closely similar triplet indices. Cluster A3 contains mainly valence/bond-corrected molecular connectivity indices. The next cluster, A4, consists mainly of the information-based indices IC (information content), SIC (structural information content), and CIC (complementary information content) for the hydrogen-filled graphs of order higher than 2 for IC and higher than 3 for SIC and CIC. Cluster A5 is composed mainly of triplet indices having as main diagonal unit vectors either distance sums or total number N of vertices.

Each of the remaining clusters has less than 10 descriptors. Clusters A6 and A7 contain mostly triplet descriptors: A6 with the distance sum S and A7 with the order N of the hydrogen-depleted graph, as the main diagonal unit vector; cluster A7 also includes two simple path cluster molecular connectivity indices. Cluster A8 contains simple cluster- and bond/valence-corrected cluster connectivities of high order (4–6). Cluster A9 again consists exclusively of triplet indices, and they are based on summing squares of LOVIs based mainly on distance sum unit vectors on the main diagonal.

Cluster A10 includes three information-theoretic indices IC and SIC of low order (0 and 1) as well as two triplet indices having in common the two unit vectors (distance sum S for the main diagonal, vertex degree V for the free term) and the operation for assembling LOVIs into an index (summation of LOVI square roots).

Interestingly, the four smallest clusters having four descriptors each are pairwise similar in type: A11 with A13, and A12 with A14. Cluster A11 consists of *information TIs* (IC, SIC, CIC) of low order (0–2), whereas A13 includes the same TIs of slightly higher order (2 and 3). Clusters A12 and A14 group together *molecular connectivity indices* based on simple cluster and simple cycle, respectively.

A general remark for the triplet indices is that what groups them together is not the matrix on which they are based (adjacency matrix or distance matrix) but the two unit vectors that convert such matrices into systems of linear equations.

Table 4. Summary of Variable Clustering for 1037 Diverse Chemicals

cluster	number of variables	representative variables (max. 25% of total listed)
B1	49	P_0, ANN_3, ANN_5, $AN1_3$, ANN_1, ANV_4, $AS1_4$, DN^21_4
B2	13	ANV_1, P_3, M_2
B3	13	$AS1_1$, $AS1_5$, $DS1_1$
B4	13	$^6\chi$, $^6\chi^b$, P_7
B5	11	ASN_5, $AS1_3$, ASN_1
B6	10	SIC_3, SIC_4, CIC_4
B7	9	$^5\chi^b{}_{PC}$, $^5\chi_{PC}$
B8	8	ASZ_2, ASZ_1
B9	6	$^5\chi^b{}_C$, $^5\chi_C$
B10	6	$^3\chi_{Ch}$, $^3\chi^b{}_{Ch}$
B11	6	IC_4, IC_5
B12	6	CIC_1, SIC_1
B13	6	$^6\chi^v{}_{Ch}$, $^6\chi^b{}_{Ch}$
B14	6	$^3\chi^b{}_C$, $^4\chi_C$
B15	4	J^B
B16	4	$AS1_2$
B17	4	DN^2N_1
B18	2	ANS_1

Clustering of Descriptors for the Diverse Set of Compounds. There are 18 variable clusters grouping together 176 variables for the database of 1037 diverse compounds (Table 4). Cluster B1, with 49 descriptors, includes 28% of all variables; 35 of these descriptors are common to cluster A1. Some of these indices, e.g., W (Wiener number), P_0 (number of non-hydrogen atoms), and P_1 (number of bonds in the hydrogen-depleted graph), express molecular size. It is interesting that most of the triplet variables (AZV_i, AZN_i, and ANN_i with $i = 1-5$ as well as several other ones) are found to be common to clusters A1 and B1. Five other descriptors ($^0\chi^b$, $^2\chi^b$, $^3\chi^b$, $^0\chi^v$, and $^3\chi^v$) also appear in both clusters A1 and B1.

Cluster B2 has nine variables in common with cluster A1; most of these ($^3\chi$, $^4\chi$, P_2-P_4) are path connectivities of intermediate order. A couple of triplet indices (ANV_1 and ANV_5) are also in common with cluster A1; another pair of triplet indices (ASN_3 and ASN_4) are in common with cluster A7.

Cluster B3 contains triplet indices with distance sums as main diagonal vector; they occur in clusters A5 and A9. In addition, two descriptors ($\bar{I}_D^W$ and H^D) appear also in cluster A1.

Cluster B4 is uniquely associated with cluster A2 and consists of indices $^5\chi$, $^6\chi$, $^5\chi^b$, $^6\chi^b$, $^5\chi^v$, $^6\chi^v$, and P_6-P_{10}. These descriptors are based on long paths; therefore, these variables appear only when large molecules are involved.

Seven of the eleven variables of cluster B5 form exclusively cluster A6; they are related to molecular shape via vertex complexity and graph radius. Five triplet indices such as ASN_1, ASN_5, DSN_1, DSN_5, and ANV_2 also are common to these two clusters.

Very interesting correspondences are manifested by cluster B6, which is mainly associated with two clusters involving the hydrocarbon database, namely, A4 and A13 (plus one descriptor in B6 that appears in A10). All variables are of information-theoretic type. These higher-order variables (SIC_3-SIC_6 and CIC_3-CIC_6) are common to clusters B6 and A4 and represent a true measure of molecular complexity. The lower- and intermediate-order indices such as IC_1 or SIC_2 that appear in clusters B6 and A10 or B6 and A13, respectively, provide information on lower-order complexity that may be more degenerate than that furnished by the higher-order information indices. One should stress here that information content indices form clusters that are separate from clusters with other descriptors, meaning that such variables convey unique information relative to structure and molecular complexity.

Cluster B7 consists only of path-cluster molecular connectivity descriptors that were included in clusters A3, A7, and A8 for the hydrocarbons.

Cluster B8 includes triplet indices, all of which have the atomic number Z for the free-term vector in the system of linear equations. Most of these descriptors appear in clusters A1, A5, and A9.

Cluster B9 consists of high-order connectivity-cluster terms all contained in cluster A8. For hydrocarbons, descriptors ${}^6\chi^b{}_C$ and ${}^6\chi^v{}_C$ are perfectly correlated with descriptor ${}^6\chi_C$; therefore, the former variables did not appear in the hydrocarbon cluster A8. For the diverse-compound database, such a correlation is not perfect because of differences in atom types.

An interesting observation concerns cluster B10: all six variables are absent from the hydrocarbon database because the database does not contain any three- or four-membered rings, unlike the diverse compound database. This is why indices ${}^{3/4}\chi_{Ch}$, ${}^{3/4}\chi^b{}_{Ch}$, and ${}^{3/4}\chi^v{}_{Ch}$ appear only in cluster B10.

Cluster B11 has all but one of its descriptors contained in cluster A4; these information content indices, IC_2-IC_6, measure a high degree of nonredundancy of topological neighborhoods.

Cluster B12 has four of its variables contained in cluster A11; these descriptors (SIC_0, CIC_0-CIC_2) express lower-order redundancy of topological neighborhoods. This is true of indices IC_0 and SIC_1 as well, which are present in cluster A10.

From cluster B13, the six descriptors (simple, bond- and valence-corrected chain molecular connectivity indices) are partitioned equally between clusters A2 and A14, according to the six- versus five-membered ring size, respectively; in the hydrocarbon database A, six-membered chain (or rings) predominate.

Cluster B14 is exclusively associated in a one-to-one relationship with cluster A12. The corresponding descriptors ${}^3\chi_C$ and ${}^4\chi_C$ as well as their bond- and valence-corrected counterparts represent connectivity indices on three- and four-vertex structural clusters. For the hydrocarbon database, we have again a case in which the two indices ${}^4\chi^b{}_C$ and ${}^4\chi^v{}_C$, perfectly correlated with ${}^4\chi_C$, do not appear explicitly in cluster A12.

Half of the variables (J-type indices) in cluster B15 are contained in cluster A2. These J indices again form a cluster apart from all other ones in the case of the diverse database, proving that when heteroatoms are taken into account, the information provided by such J-type indices is unique.

Clusters B16, B17, and B18 each have a small number of triplet-type descriptors; the three descriptors of cluster B17 are all contained in cluster A7.

Intercluster Correlations. From each cluster we select 15–25% of the descriptors according to the maximal value of the correlation coefficient with their own cluster. In most cases, the first selected descriptor also has the minimal value of the correlation with the next closest cluster, expressed by the $1 - r^2$ value. When more than one index is chosen from the same cluster, after the first one was selected as indicated above, the next one must also fulfill a third criterion, namely, a low intercorrelation with the previously selected indices of the same cluster.

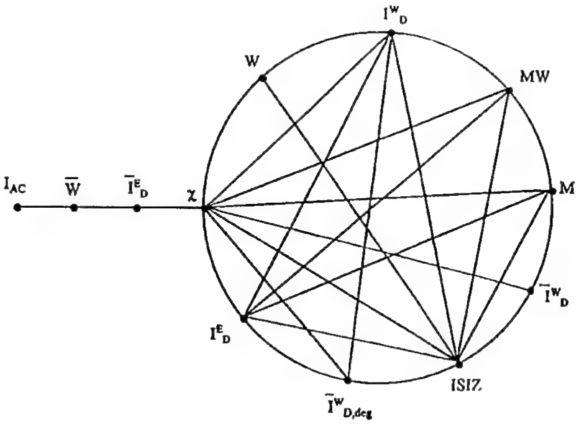


Figure 2. Graph of highly correlated topological indices (TIs) according to Todeschini et al. (notation of TIs as in Table 3 of ref 31). Lines connect TIs with $r > 0.90$.

There were four intercluster correlations within the hydrocarbon data set that were greater than 0.9, and all involved cluster A1. Cluster A1 was positively correlated with A2, A3, and A7. Cluster A1 was correlated negatively with A5. Each of the clusters characterizes some aspect of molecular size and shape.

Cluster B1 showed an intercluster correlation of 0.92 with cluster B2 and −0.90 with cluster B3. These were the only intercluster correlations greater than 0.9. These clusters are the three largest clusters in set B. Like cluster A1, cluster B1 groups TIs expressing molecular size and shape. Interestingly, in set A cluster A1 also had a negative intercluster correlation with cluster A5; it is therefore not surprising that clusters A5 and B3 have the most abundantly populated line connecting them in Figure 1.

In summary, for the hydrocarbon database there are four intercluster correlations with $r > 0.90$ all involving on one hand the first cluster A1 and on the other hand clusters A2, A3, A5, and A7. For the diverse compound database there are only two such intercluster correlations with $r > 0.90$, namely, B1 with B2 and B3. This is not unexpected, as the combination of the first three clusters in each case contains more descriptors than the parameters remaining in all the remaining ones together.

In this context, one should mention that Todeschini and co-workers published an interesting study[35] on 23 TIs for a set of 667 diverse chemicals, 20% of which were hydrocarbons; the above authors excluded 10 of these TIs because they were degenerate, or redundant or had intercorrelation factors higher than 0.90. A graph depicting highly intercorrelated indices using data published by these authors is presented in Figure 2, which is similar to a graph published earlier.[22]

Ten TIs were then selected by Todeschini et al.,[35] namely, the molecular weight (M_W), J, IC, CIC, the bonding information content (BIC), mean Randić connectivity (χ), the information content on atomic composition (I_{AC}), the mean Wiener index ($\bar{W}$), and the mean information indices on equality of distance degree and on the magnitude of distance degree ($\bar{I}^{E}_{D,deg}$ and $\bar{I}^{W}_{D,deg}$, respectively). Then, using principal-component analysis for the above 10 TIs, Todeschini et al. analyzed the composition of the first six principal-components. They found that the first PC is mainly composed of indices that express the size of molecules (M_W, $\bar{W}$, IC, $\bar{I}^{E}_{D,deg}$ and $\bar{I}^{W}_{D,deg}$). This is in agreement with the earlier finding of Basak et al. for a set of 3692 diverse chemicals that the first PC is related to molecular size.[29] Further, Todeschini et al. found that the second PC is dominated by indices expressing information on bonds (IC, CIC, and BIC). This is also analogous to the results reported by Basak et al.[29] that the second axis represents molecular complexity as encoded by higher-order neighborhood complexity indices (IC_2, IC_3, SIC_2, SIC_3, CIC_2, CIC_3, etc.). The IC, CIC, and BIC indices used by Todeschini et al. are based solely on first-order topological bonding/neighborhoods and slightly different equivalence relations as compared to the IC_r, SIC_r, and CIC_r indices defined by Roy et al.[27] In studies by Basak et al.,[29] the first-order complexity indices (IC_1, SIC_1, CIC_1) were usually most correlated with the first PC. Each of the next four PCs in Todeschini et al.'s study[35] is dominated by a single TI,viz., χ, I_{AC}, J (indicating branching), and $\bar{I}^{E}_{D,deg}$ (connected with the position of substituents on the molecular scaffold), respectively.

ACKNOWLEDGMENT

This is contribution number 251 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by Grants F49620-94-1-0401 and F49620-96-1-0330 from the U.S. Air Force.

REFERENCES AND NOTES

(1) Devillers, J., Balaban, A. T., Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999. (a) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. Chapter 2. (b) Kier, L. H.; Hall, L. B. Molecular Connectivity Chi Indices for Database Analysis and Structure−Property Modeling. Chapter 7. (c) Kier, L. H.; Hall, L. B. The Kappa Indices for Molecular Modeling of Molecular Shape and Flexibility. Chapter 10. (d) Kier, L. H.; Hall, L. B. The Electrotopological State: Structure Modeling for QSAR and Database Analysis. Chapter 11. (e) Basak, S. C. Information-Theoretic Indices of Neighborhood Complexity and Their Application. Chapter 12. (f) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Hierarchical Approach to the Development of QSAR Models Using Topological, Geometrical and Quantum Chemical Parameters. Chapter 14.

(2) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(3) (a) Balaban, A. T. Chemical Graphs. Part 35. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *5*, 239−261. (b) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Centre Concept and Derived Topological Indices. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 196−213. (c) Balaban, A. T.; Bertelsen, S.; Basak, S. C. New Centric Topological Indexes for Acyclic Molecules (Trees) and Substituents (Rooted Trees), and Coding of Rooted Trees, *Math. Chem. (MATCH)* **1994**, *30*, 55−72.

(4) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(5) (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976. (b) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Studies*; Research Studies Press: Letchworth, 1986. (c) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.

(6) (a) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *80*, 399−404. (b) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206. (c) Balaban, A. T. Chemical Graphs. 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122. (d) Balaban, A. T.; Filip, P. Computer Program for Topological Index J (Average Distance Sum Connectivity). *Math. Chem. (MATCH)* **1984**, *16*, 163−190.

(7) Bonchev, D. *Information-Theoretic Indices for Characterization of Chemical Structure*; Research Studies Press: Letchworth, 1993.

(8) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992; pp 225−273.

(9) Balaban, A. T. Using Real Numbers as Vertex Invariants for Third-Generation Topological Indices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23−28.

(10) Basak, S. C.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological Indices: Their Nature, Mutual Relatedness, and Applications. *Math. Modell.* **1987**, *8*, 300−305.

(11) Gutman, I.; Ruscic, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.

(12) Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.

(13) Balaban, A. T.; Balaban, T. S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383−397.

(14) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61−83.

(15) (a) Ivanciuc, O.; Balaban, T. S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318. (b) Plavsic, D.; Nikolić, S.; Trinajstić, N.; Mihalic, Z. On the Harary Index for Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235−250.

(16) (a) Balaban, A. T.; Diudea, M. V. Real Number Vertex Invariants: Regressive Distance Sums and Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 421−428. (b) Diudea, M. V.; Minailiuc, O.; Balaban, A. T. Molecular Topology. Part 4. Regessive Vertex Degrees (New Graph Invariants) and Derived Topological Indices. *J. Comput. Chem.* **1991**, *12*, 527−535.

(17) Diudea, M. V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged Matrices and Related Numbers. *Math. Chem. (MATCH)* **1997**, *35*, 129−143.

(18) Klein, D. J.; Randić, M. Resistance Distance. *J. Math. Chem.* **1993**, *17*, 147−154.

(19) (a) Randić, M.; Basak, S. C. Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261−266. (b) Basak, S. C.; Gute, B. D. Characterization of Molecular Structures Using Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1−21. (c) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131. (d) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272. (e) Basak, S. C.; Grunwald, G. D. Use of Graph Invariants, Volume and Total Surface Area in Predicting Boiling Point of Alkanes. *Math. Modell. Sci. Comput.* **1993**, *2*, 735−740. (f) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modell. Sci. Comput.*, in press. (g) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Modell. Comput. Sci.*, in press. (h) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modell. Sci. Comput.*, in press. (i) Basak, S. C.; Grunwald, G. D. Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **1995**, *31*, 2529−2546. (j) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett.* **1995**, *79*, 239−250. (k) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: A Molecular Similarity Approach. In *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*; Johnson, B. L., Xintaras, C., Andrews, J. S., Eds.; Princeton Scientific Publishing Co. Inc.: Princeton, NJ, 1997; pp 492−504. (l) Basak, S. C.; Grunwald, G. D. Predicting Genotoxicity of Chemicals Using Nonempirical Parameters. In *Proceeding of XVI International Cancer Congress*; R. S. Rao, M. G. Deo, L. D. Sanghvi, Eds.; Monduzzi Editore S.p.A.: Bologna, Italy, 1995; pp 413−416. (m) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical and Neural Network Methods for Predicting Toxic Modes of Action of Chemicals, *Environ. Toxicol. Chem.* **1998**, *17*, 1056−1064. (n) Basak, S. C.; Veith, G. D.; Grunwald, G. D. Prediction of Octanol−Water Partition Coefficient (K_{ow}) Using

Algorithmically-Derived Variables G J. *Environ. Toxicol. Chem.* **1992**, *11*, 893−900. (o) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm. Res.* **1996**, *13*, 775−778. (p) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol−Water Partition Coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060. (q) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.

(20) Ivanciuc, O.; Balaban, A. T. Investigation of Alkane Branching with Topological Indices. *Math. Chem. (MATCH)*, in press.

(21) Balaban, A. T., Ed. *From Chemical Topology to Three-Dimensional Geometry*; Plenum Press: New York, 1998.

(22) Motoc, I.; Balaban, A. T. Topological Indices: Intercorrelations, Physical Meaning, Correlational Ability. *Rev. Roum. Chim.* **1981**, *26*, 593−600. Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological Indices: Inter-Relations and Composition. *Math. Chem. (MATCH)* **1982**, *13*, 369−404.

(23) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indices for Structure−Activity Correlations. In *Steric Effects in Drug Design*; Charton, M., Motoc, I., Eds.; *Top. Curr. Chem.* **1983**, *114*, 21−55. Balaban, A. T.; Niculescu-Duvaz, I.;. Simon, Z. Topological Aspects in QSAR for Biologically-Active Molecules. *Acta Pharm. Jugosl.* 1987, *37*, 7−36. Voiculetz, N.; Balaban, A. T. Niculescu-Duvaz, I.; Simon, Z. *Modeling of Cancer Genesis and Prevention*; CRC Press: Boca Raton, FL, 1990.

(24) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the Molecular Similarity of Chemicals Using Topological Indices. In *Advances in Molecular Similarity*, Vol. 2; R. Carbo-Dorca, P. G. Mezey, Eds.; JAI Press: Stanford, CT, 1998; pp 171−185.

(25) Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular Modelling of the Physical Properties of Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194. Mekenyan, O.; Bonchev, D.; Trinajstić, N. Chemical Graph Theory: Modelling the Thermodynamic Properties of Molecules. *Int. J. Quantum Chem.* 1980, *18*, 369−380. Karcher, W. *Spectral Atlas of Polycyclic Aromatic Hydrocarbons*; Kluwer Academic Press: Dordrecht, 1988; Vol. 2. pp 16−19.

(26) Russom, C. L. *Assessment Tools for the Evaluation of Risk (ASTER)*, v. 1.0; U.S. Environmental Protection Agency, 1992.

(27) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; 4th International Conference, Zurich; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York; 1983; pp 745−750.

(28) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609. Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. A Quantitative Structure Activity Relationship Study of Tumor Inhibitory Triazenes Using Bonding Information Content and Lipophilicity. *ICRS Med. Sci.* **1982**, *10*, 933−934. Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. A Quantitative Structure−Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneimitt.-Forsch. Drug Res.* **1983**, *33*, 501−503.

(29) Basak, S. C.; Magnuson, V. R. Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discr. Appl. Math.* 1988, *19*, 17−44.

(30) Balasubramanian, K.; Basak, S. C. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367−373.

(31) (a) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modell. Sci. Comput.*, in press. (b) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modell. Sci. Comput.*, in press. (c) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Modell. Sci. Comput.*, in press.

(32) Raychaudhury, C.; Ray, S. K.; Roy, A. B.; Ghosh, J. J.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Indices. *J. Comput. Chem.* **1984**, *5*, 581−588.

(33) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY 2.3*: University of Minnesota, 1988.

(34) SAS Institute Inc. The VARCLUS Procedure. In *SAS/STAT User's Guide*, Version 6, fourth ed.; SAS Institute Inc.: Cary, NC, 1989; Vol. 2, 846 pp.

(35) Todeschini, R.; Cazar, R.; Collina, E. The Chemical Meaning of Topological Indices. *Chemometrics Intell. Lab. Syst.* 1992, *15*, 51−59.

CI990114Y

APPENDIX 1.13 Use of graph invariants in QMSA and predictive toxicology

DIMACS Series in Discrete Mathematics
and Theoretical Computer Science
Volume 51, 2000

Use of Graph Invariants in QMSA and Predictive Toxicology

S.C. Basak and B.D. Gute

ABSTRACT. Mathematical invariants are frequently used for the characterization of molecular graphs. Such invariants quantify structural features of chemicals like size, shape, symmetry, cyclicity, complexity, branching, etc. Numerical graph invariants or topological indices (TIs) have been used in developing quantitative structure-property/ activity/ toxicity relationship models and in defining intermolecular similarity. In this paper, we have used a set of TIs and a class of substructures called atom pairs (APs) in selecting analogs of probe chemicals from a set of mutagens. The result shows that both of the similarity methods select analogs which have reasonable structural similarity with the query chemicals. Such analogs, selected computationally, can be useful in the hazard assessment of chemicals for which very little or no toxicity data are available.

1. Introduction

A contemporary interest in mathematical chemistry is the characterization of molecular structure using graph theoretic formalism [1]–[11]. A graph $G = [V, E]$ consists of an ordered pair of two sets V and E, representing the vertices and edges, respectively. G becomes a molecular graph when the set V represents the set of atoms in a molecule and the set E symbolizes chemical bonds between adjacent atoms [8].

Mathematical characterization of molecular graphs (structures) may be accomplished using graph invariants. An invariant may be a polynomial, a sequence of numbers, or a real number. A real number characterizing a molecular graph is called a topological index (TI). TIs quantify different aspects of molecular architecture, viz., size, shape, cyclicity, branching, symmetry, etc [8].

TIs have been used extensively in quantitative structure-property/activity relationships (QSPR and QSAR respectively) and the quantification of intermolecular similarity/dissimilarity of chemicals [10]–[24]. In quantitative molecular similarity analysis (QMSA) studies, TIs have been used to derive high dimensional structure spaces where the Euclidean distance D_{ij} between a pair of molecules i and j is used

1991 *Mathematics Subject Classification.* Primary 54C40, 14E20; Secondary 46E25, 20C20.

This is contribution number 239 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force. The authors would like to extend their thanks to Greg Grunwald for technical support.

to quantify the similarity between them. Similarity measures can be used either for the selection of analogs of chemicals or in the prediction of the property/activity of a molecule from the property of its selected neighbor(s).

In some of our recent QSAR/QMSA studies we have used different similarity measures derived from TIs in the selection of analogs and prediction of properties/activities for diverse sets of chemicals. We have also used orthogonal descriptors derived from a set of over 100 graph invariants to estimate bioactivity/toxicity of different graphs of molecules. In this paper we have used similarity measures derived from TIs in: a) selecting analogs of an isospectral graph from a diverse set of 221 compounds, and b) predicting the mutagenicity of a set of 113 mutagens and non-mutagens using QMSA methods.

2. Methods

2.1. Databases. A set of 19 pairs of isospectral graphs from the work of Balasubramanian and Basak [25] were added to a set of 107 benzamidines [26] and a composite set of 76 diverse compounds used in an earlier study by Basak and Grunwald [23] to create a varied library of 221 compounds. This composite library was created to provide a large set containing both congeneric and non-congeneric sets to test analog selection methods. The chemical structures for the 19 pairs of isospectral graphs have been presented previously [25].

A second data set, representing a subset of the set of 277 chemicals presented by Yamaguchi *et al.* [27] was also used in the current study. This subset consisted of all the chemicals in the set of 277 chemicals that had reported results for mutagenicity in the Ames test, mutagenicity in the medium term liver carcinogenesis bioassay, and carcinogenicity in the two-year rodent bioassay in rat and/or mouse. This subseting resulted in a set of 113 chemicals, 68 of which are classified as non-mutagens and 45 of which are classified as mutagens in the Ames test. This set of chemicals and their observed mutagenicity are reported in Table 1.

TABLE 1: Mutagenicity in the Ames test for 113 chemicals

No.[a]	Compound Name	Obs. Ames Mutagenicity
1.5	butylated hydroxyanisole (BHA)	0
1.6	caffeic acid	0
1.7	catechol	0
1.8	clofibrate	0
1.9	di(2-ethylhexyl)phthalate (DEHP)	0
1.10	hydroquinone	0
1.11	p-methoxyphenol	0
1.12	sesamol	0
1.13	tamoxifen	0
1.14	acetaminophen	0
1.15	benzoin	0
1.16	EPN	0
1.17	gallic acid	0
1.18	a-tocopherol	0
2.2	2-acethylaminofluorene (AAF)	1

TABLE 1: Mutagenicity in the Ames test for 113 chemicals

No.[a]	Compound Name	Obs. Ames Mutagenicity
2.3	adriamycin	1
2.4	aflatoxin B1	1
2.5	benzo[a]pyrene	1
2.7	captafol	1
2.8	captan	1
2.9	carbazole	1
2.10	dibutylnitrosamine (DBN)	1
2.11	diethylnitrosamine (DEN)	1
2.12	3,2'-dimethyl-4-aminobiphenyl (DMAB)	1
2.14	dimethylnitrosamine (DMN)	1
2.15	N-ethyl-N-hydroxyethylnitrosamine (EHEN)	1
2.16	N-ethyl-N-nitrosourea (ENU)	1
2.20	hydrazobenzene	1
2.22	laciocarpine	1
2.26	3'-methyl-4-dimethylaminoazobenzene (3'-Me-DAB)	1
2.27	3-amino-9-ethylcarbazole	1
2.28	N-nitrosooxazolidine	1
2.29	N-nitrosodi-n-propylamine (NDPA)	1
2.30	N-nitrosomorpholine	1
2.31	N-nitrosopiperidine	1
2.32	N-nitrosopyrrolidine	1
2.33	quinoline	1
2.34	sterigmatocystin	1
2.35	4,4'-thiodianiline	1
2.42	alachlor	0
2.43	aldrin	0
2.44	auramine O	0
2.45	barbital	0
2.46	chlordane	0
2.47	chlorendic acid	0
2.48	chlorobenzilate	0
2.49	DDT	0
2.50	dieldrin	0
2.51	diethylstilbestrol	0
2.53	ethenzamide	0
2.54	17α-ethinyl estradiol	0
2.55	DL-ethionine	0
2.56	hexachlorobenzene (HCB)	0
2.57	a-hexachlorocyclohexane (a-HCH)	0
2.58	d-limonene	0
2.59	monoclotaline	0
2.60	N-nitrosodiethanolamine	0
2.61	phenobarbital	0
2.64	safrole	0

TABLE 1: Mutagenicity in the Ames test for 113 chemicals

No.[a]	Compound Name	Obs. Ames Mutagenicity
2.66	thioacetamide	0
2.67	triadimefon	0
2.68	trifluralin	0
2.69	urethane	0
2.70	polychlorinated biphenyl (PCB)	0
2.71	malathion	0
2.72	vinclozolin	0
3.1	acetophenetidine (phenacetin)	1
3.2	azathioprine	1
3.3	N-butyl-N-(4-hydroxybutyl)nitrosamine (BBN)	1
3.4	chrysazin (danthron)	1
3.5	4,4'-diaminodiphenylmethane (DDPM)	1
3.6	7,12-dimethylbenz[a]anthracene (DMBA)	1
3.7	N-ethyl-N-(4-hydroxybutyl)nitrosamine (EHBN)	1
3.8	folpet	1
3.9	hydrogen peroxide	1
3.11	3-methylcholanthrene (3-MC)	1
3.12	N-methyl-N'-nitro-N-nitrosoguanidine (MNNG)	1
3.13	N-methyl-N-nitrosourea (MNU)	1
3.14	8-nitroquinoline	1
3.17	streptozotocin	1
3.18	o-toluidine	1
3.20	6-methylquinoline	1
3.21	8-methylquinoline	1
3.22	nitrofrantoln	1
3.23	6-nitroquinoline	1
3.24	quercetin	1
3.32	acetaldehyde	0
3.33	atrazine	0
3.34	di(2-ethylhexyl)adipate (DEHA)	0
3.35	1,1-dimethylhydrazine	0
3.39	trichloroacetic acid	0
3.42	4-acethylaminofluorene (AAF)	0
3.43	aspirin	0
3.44	butylated hydroxytoluene (BHT)	0
3.45	caffeine	0
3.46	caprolactam	0
3.47	chenodeoxicholic acid	0
3.49	cypermethrin	0
3.50	deltamethrin	0
3.51	diltiazem	0
3.52	dimethylsulfoxide (DMSO)	0
3.53	diazinon	0
3.54	fenvalerate	0

TABLE 1: Mutagenicity in the Ames test for 113 chemicals

No.[a]	Compound Name	Obs. Ames Mutagenicity
3.55	glutathione	0
3.56	4-o-hexyl-2,3,6-trimethylhydroquinone (HTHQ)	0
3.58	lithocolic acid	0
3.59	d-mannitol	0
3.61	phenol	0
3.64	propyl galiate	0
3.65	propylparaben	0
3.66	pyrene	0
3.67	resorcinol	0
3.71	trimorphamide	0

[a] The numbering scheme refers to the enumeration of the chemicals in the presentation by Yamaguchi *et al.* [27] where the numeral before the decimal place refers to the table in which the compound was listed (see below) and the numerals after the decimal refer to the compounds location within the table.
Table 1 - Association between inhibitory results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.
Table 2 - Association between positive results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.
Table 3 - Association between negative results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.

2.2. Calculation of Topological Indices. The TIs calculated for this study are listed in Table 2 and include Wiener number [28], molecular connectivity indices as calculated by Randić [29] and Kier and Hall [4], frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [30] as well as those of Raychaudhury *et al.* [31], parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [32]–[34], and Balaban's J indices [35]–[37]. The majority of the TIs were calculated using POLLY 2.3 [38]. The J indices were calculated using software developed by the authors.

The Wiener index (W) [28], the first topological index reported in the chemical literature, may be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric $n \times n$ matrix (d_{ij}), where d_{ij} is equal to the distance between vertices v_i and v_j in G. Each diagonal element d_{ii} of D(G) is zero. We give below the distance matrix $D(G_1)$ of the unlabeled hydrogen-suppressed graph G_1 of thioacetamide (Fig. 1):

$$D(G_1) = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 & 2 \\ 4 & 2 & 1 & 2 & 0 \end{array}$$

W is calculated as:

$$W = 1/2 \sum_{ij} d_{ij} = \sum_{h} h \cdot g_h \tag{2.1}$$

where g_h is the number of unordered pairs of vertices whose distance is h. Thus for $D(G_1)$, W has a value of nine.

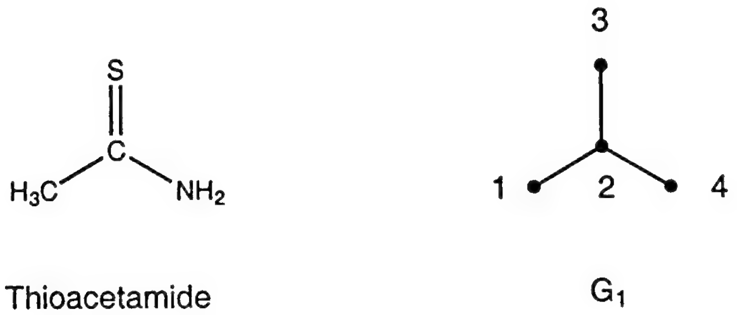


FIGURE 1. Unlabeled, hydrogen-suppressed graph of thioacetamide (G_1)

Randić's connectivity index [29], and higher-order connectivity path, cluster, path-cluster and chain types of simple, bond and valence connectivity parameters were calculated using the method of Kier and Hall [4]. The generalized form of the simple path connectivity index is as follows:

$$^{h}\chi = \sum (v_i v_j \ldots v_{h+1})^{-1/2} \tag{2.2}$$

where $v_i, v_j, \ldots, v_{h+1}$ are the degrees of the vertices in the path of length h. The path length parameters (P_h), number of paths of length h ($h = 0, 1, \ldots, 10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set A of n elements is derived from a molecular graph G depending upon certain structural characteristics. On the basis of an equivalence relation defined on A, the set A is partitioned into h disjoint subsets A_i of order $n_i (i = 1, 2, \ldots, h; \sum_{i=1}^{h} n_i = n)$. A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \ldots, A_h$$
$$p_1, p_2, \ldots, p_h$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the i^{th} subset.

TABLE 2: Symbols and brief definitions for 101 topological indices

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0 - 6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0 - 6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0 - 6$) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi$	Path connectivity index of order $h = 0 - 6$
${}^h\chi_C$	Cluster connectivity index of order $h = 3 - 6$
${}^h\chi_{Ch}$	Chain connectivity index of order $h = 3 - 6$
${}^h\chi_{PC}$	Path-cluster connectivity index of order $h = 4 - 6$
${}^h\chi^b$	Bond path connectivity index of order $h = 0 - 6$
${}^h\chi^b_C$	Bond cluster connectivity index of order $h = 3 - 6$
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order $h = 3 - 6$
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order $h = 4 - 6$
${}^h\chi^v$	Valence path connectivity index of order $h = 0 - 6$
${}^h\chi^v_C$	Valence cluster connectivity index of order $h = 3 - 6$
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order $h = 3 - 6$
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order $h = 4 - 6$
P_h	Number of paths of length $h = 0 - 10$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

The mean information content of an element of A is defined by Shannon's relation [39]:

$$IC = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{2.3}$$

The logarithm is taken at base 2 for measuring the information content in bits. The total information content of the set A is then $n \times IC$. Figure 2 provides a sample calculation for IC_1.

It is to be noted that the information content of a graph G is not uniquely defined. It depends on how the set A is derived from G as well as on the equivalence relation which partitions A into disjoint subsets A_i. For example, when A constitutes the vertex set of a chemical graph G, two methods of partitioning have been widely used: a) chromatic-number coloring of G where two vertices of the same color are considered equivalent, and b) determination of the orbits of the automorphism group of G thereafter vertices belonging to the same orbit are considered equivalent.

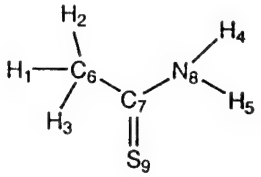
Rashevsky was the first to calculate the information content of graphs where "topologically equivalent" vertices were placed in the same equivalence class [40]. In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex $u_i (i = 1, 2, \ldots, k)$ of the vertex u, there is a distinct neighboring vertex v_i of the same degree for the vertex v. While Rashevsky used simple linear graphs with indistinguishable vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, i.e., electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* [41] calculated information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by stereo-electronic characteristics of distant neighbors, i.e., neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any non-negative real number and v is a vertex of the graph G, then the open sphere $S(v,r)$ is defined as the set consisting of all vertices v_i in G such that $d(v, v_i) < r$. Therefore, $S(v,0) = \emptyset$, $S(v,r) = v$ for $0 < r < 1$, and $S(v,r)$ is the set consisting of v and all vertices v_i of G situated at unit distance from v, if $1 < r < 2$.

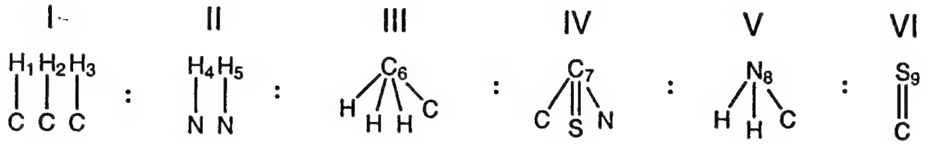
One can construct such open spheres for higher integral values of r. For a particular value of r, the collection of all such open spheres $S(v,r)$, where v runs over the whole vertex set V, forms a neighborhood system of the vertices of G. A suitably defined equivalence relation can then partition V into disjoint subsets consisting of vertices which are topologically equivalent for r^{th} order neighborhood. Such an approach has been developed and the information-theoretic indices calculated based on this idea are called indices of neighborhood symmetry [34].

In this method, chemicals are symbolized by weighted linear graphs. Two vertices u_o and v_o of a molecular graph are said to be equivalent with respect to r^{th} order neighborhood if and only if corresponding to each path $u_o, u_1, \ldots, u_r$ of

G_2: thioacetamide



First order neighbors:



Subsets:

I	II	III	IV	V	VI
(H_1-H_3)	(H_4-H_5)	C_6	C_7	N_8	S_9

Probability:

I	II	III	IV	V	VI
3/9	2/9	1/9	1/9	1/9	1/9

$$IC_1 = 4 * 1/9 * \mathrm{Log}_2\ 9 + 2/9 * \mathrm{Log}_2\ 9/2 + 3/9 * \mathrm{Log}_2\ 9/3 \qquad = 2.419 \text{ bits}$$

$$SIC_1 = IC_1/\mathrm{Log}_2\ 9 \qquad = 0.763 \text{ bits}$$

$$CIC_1 = \mathrm{Log}_2\ 12 - IC_2 \qquad = 0.751 \text{ bits}$$

FIGURE 2. Labeled, hydrogen-filled graph of thioacetamide (G_2) and sample calculations for IC_1, SIC_1 and CIC_1

length r, there is a distinct path $v_o, v1, \ldots, v_r$ of the same length such that the paths have similar edge weights, and both u_o and v_o are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation has been described in earlier studies [**34, 42**].

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by Eq. (2.2). Basak *et al.* [32] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r / \log_2 n \tag{2.4}$$

where IC_r is calculated from Eq.(2.2) and n is the total number of vertices of the graph.

Another information-theoretic invariant, complementary information content (CIC_r) [43], is defined as:

$$CIC_r = \log_2 n - IC_r \tag{2.5}$$

CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r. Sample calculations for SIC_1 and CIC_1 have been included in Figure 2.

The information-theoretic index on graph distance, I_D^W is calculated from the distance matrix $D(G)$ of a chemical graph G as follows [30]:

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \tag{2.6}$$

The mean information index, $\bar{I}_D^W$, is found by dividing the information index I_D^W by W. The information theoretic parameters defined on the distance matrix, H^D and H^V, were calculated by the method of Raychaudhury *et al.* [31].

Balaban defined a series of indices based upon distance sums within the distance matrix for a chemical graph that he designated as J indices [35]–[37]. These indices are highly discriminating with low degeneracy. Unlike W, the J indices range of values are independent of molecular size. The general form of the J index calculation is as follows:

$$J = q(\mu + 1)^{-1} \sum_{i,j,edges} (s_i s_j)^{-1/2} \tag{2.7}$$

where the cyclomatic number μ (or number of rings in the graph) is $\mu = q - n + 1$, with q edges and n vertices and s_i is the sum of the distances of atom i to all other atoms and s_j is the sum of the distances of atom j to all other atoms [35]. Variants were proposed by Balaban for incorporating information on bond type, relative electronegativities, and relative covalent radii [36, 37].

2.3. Calculation of Atom Pairs. Atom pairs (APs) were calculated using the method of Carhart *et al.* [3]. An atom pair is defined as a substructure consisting of two non-hydrogen atoms i and j and their interatomic separation:

$$< \text{atom descriptor}_i > - < \text{separation} > - < \text{atom descriptor}_j >$$

where < atom descriptor > contains information about the atomic type, number of non-hydrogen neighbors and the number of π electrons. The interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. APs used in this study were calculated by the APProbe software [43].

2.4. Statistical Methods and Computation of Intermolecular Similarity.

2.4.1. *Data Reduction.* Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some TIs may be several orders of magnitude greater than other TIs.

A principal component analysis (PCA) was used on the transformed indices to minimize the intercorrelation of indices. The PCA was conducted using the SAS procedure PRINCOMP [**44**]. The PCA produces linear combinations of the TIs, called principal components (PCs) which are derived from the correlation matrix. The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to the previous PCs, eliminating any redundancies that could occur within the set of TIs. The maximum number of PCs generated is equal to the number of TIs available. For the purposes of this study, only PCs with eigenvalues greater than one were retained. A more detailed explanation of this approach has been provided in a previous study by Basak *et al.*[**13**]. These PCs were subsequently used to determine similarity scores as described below.

2.4.2. *Similarity Measures.* Intermolecular similarity was measured using two distinct methods. The AP method uses an associative measure described by Carhart *et al.* [**3**] and is based on atom pair descriptors. The measurement is the ratio of the number of shared atom pairs between two molecules over the total number of atom pairs present in the two molecules. Similarity (S) between molecules i and j is defined as:

$$S_{ij} = 2C/(T_i + T_j) \tag{2.8}$$

where C is the number of atom pairs common to molecule i and j. T_i and T_j are the total number of atom pairs in molecule i and j, respectively. The numerator is multiplied by a factor of 2 to reflect the presence of shared atom pairs in both compounds.

The second similarity method, Euclidean distance (ED) within an n-dimensional PC space derived from TIs was used. ED between molecules i and j is defined as:

$$ED_{ij} = \left[\sum_{k=1}^{n}(D_{ik} - D_{jk})^2\right]^{1/2} \tag{2.9}$$

where n equals the number of dimensions or PCs retained from the PCA. D_{ik} and D_{jk} are the data values of the k^{th} dimension for molecules i and j, respectively.

2.4.3. *Analog / K-Nearest Neighbor Selection.* Following the quantification of intermolecular similarity of the molecules, analogs or nearest neighbors are determined on the basis of both S and ED. In the case of the AP method, two molecules are considered identical if $S = 1$, while they have no atom pairs in common if $S = 0$. The ED method measures a distance between molecules, thus the lower the value of ED the greater the similarity between two molecules.

2.4.4. *Property Estimation.* Since the data presented in the work of Yamaguchi *et al.* [**27**] represented mutagenicity as non-mutagen (−) or mutagen (+) this data was treated as a zero-one relationship, where non-mutagens have a value of zero and mutagens have a value of one. In estimating the mutagenicity of the probe compound, the mean of the observed mutagenicity of the K-nearest neighbors was used as the estimate. Thus, if the mean resulted in a value greater than 0.5, the compound was classified as a mutagen. However, if the mean was equal to 0.5, the compound was not classified as the results were inconclusive.

3. Results

3.1. Principal Component Analysis. From the PCA of the 102 TIs, eight PCs with eigenvalues greater than one were retained. These eight PCs explained, cumulatively, 95.2% of the total variance within the TI data. Table 3 lists the eigenvalues of the eight PCs, the proportion of variance explained by each PC, the cumulative variance explained, and the two TIs most correlated with each individual PC.

TABLE 3. Eigenvalues, variance explained and two TIs most correlated with the eight principal components

		Percent	Cumulative		
		variance	**variance**	**First most**	**Second most**
PC	**Eigenvalue**	**explained**	**explained**	**correlated TI**	**correlated TI**
PC_1	55.52	54.97	54.97	${}^4\chi^b$ (96.5%)	${}^3\chi$ (96.4%)
PC_2	12.38	12.26	67.23	SIC_3 (86.4%)	SIC_4 (85.5%)
PC_3	11.73	11.61	78.84	${}^5\chi^b_{Ch}$ (77.3%)	${}^5\chi^v_{Ch}$ (76.1%)
PC_4	6.78	6.71	85.55	IC_0 (55.0%)	${}^4\chi^v_{Ch}$ (49.7%)
PC_5	4.60	4.55	90.10	J (68.9%)	J^Y (62.4%)
PC_6	2.35	2.32	92.43	IC_0 (−47.2%)	SIC_0 (−36.4%)
PC_7	1.65	1.63	94.06	${}^4\chi^b_C$ (44.4%)	${}^4\chi^v_C$ (43.5%)
PC_8	1.16	1.14	95.21	${}^4\chi^v_C$ (−34.6%)	${}^6\chi^b_C$ (23.0%)

3.2. Analog Selection. Figure 3 shows the results of the analog selection for isospectral graph 10.1.1 using atom pairs to derive a similarity space and PCs to derive a Euclidean distance space. The first five analogs (neighbors) for the probe compound, 10.1.1, are presented for each of the similarity methods.

3.3. *K*-Nearest Neighbor Estimation. Table 4 presents the results for the prediction of mutagenicity for the 113 molecules over a range of K values ($K = 1-5$) for both the AP and ED methods. The results are presented as percent correctly classified and over-all percent correct prediction rates are provided as a means of comparing the efficacy of the individual models. The variability between the K levels is easily explained by the problematic nature of using a binary relationship such as this one in estimation. When the number of neighbors was even, the potential for unclassified compounds led to lower prediction rates than in the case of an odd number of neighbors.

4. Discussion

The major objective of this paper was to study the effectiveness of mathematical invariants in the characterization of molecular structure and the estimation of the toxicity of chemicals. An invariant maps a chemical structure into the set R of real numbers. A specific invariant may be used for the ordering or partial ordering of sets of molecules or in structure-activity relationship studies [45]. A particular

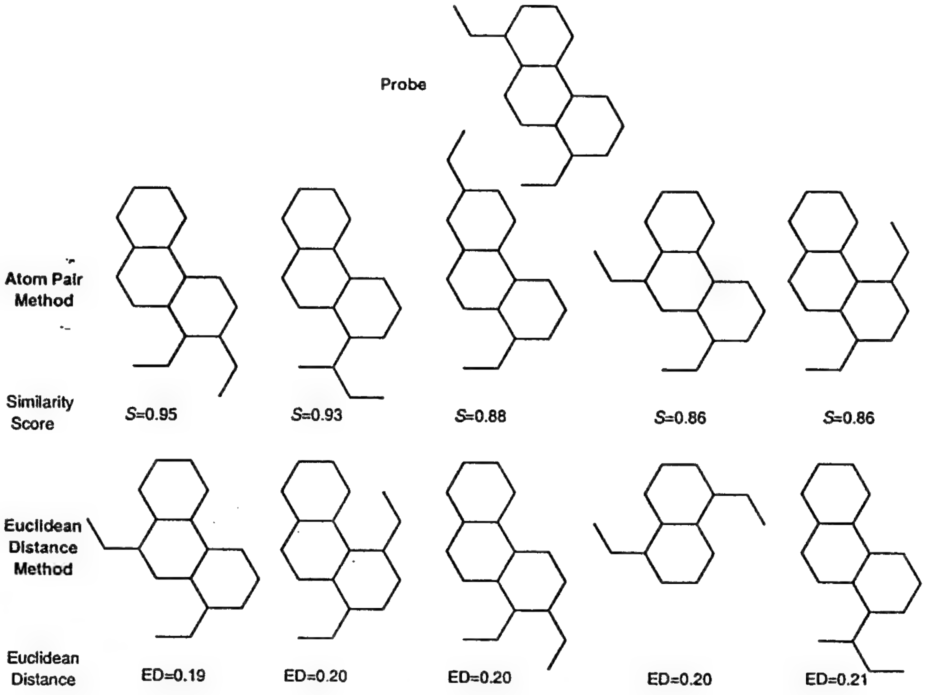


FIGURE 3. Analogs selected for isospectral graph 10.1.1

TABLE 4. KNN results for the prediction of mutagenicity for 113 chemicals

	Percent Negative Correct		Percent Positive Correct		Total Percent Correct	
K	*AP*	*ED*	*AP*	*ED*	*AP*	*ED*
1	73.5	75.0	84.1	66.7	77.7	71.7
2	66.2	64.7	72.7	33.3	68.8	52.2
3	77.9	80.9	88.6	53.3	82.1	69.9
4	70.6	69.1	77.3	42.2	73.2	58.4
5	79.4	77.9	86.4	53.3	82.1	68.1

structural invariant quantifies distinct aspects of molecular structure. Therefore, a combination of such indices might be more powerful in the mathematical characterization of molecular structure as compared to the use of one specific invariant. The problem arises out of the fact that often the various graph theoretic indices of molecular structures are strongly correlated. We have attempted to resolve this problem through the implementation of a PCA to derive orthogonal variables from a large set of calculated TIs, and using the orthogonal parameters in the characterization of structure [**10**, **12**, **15**, **17**, **18**, **22**, **23**].

In the present study we have used calculated atom pairs and principal components derived from TIs to select structural analogs for a probe compound from a diverse set containing closely related structures. The result of this analog selection, depicted in Figure 3, shows that the five neighbors selected by each of the methods exhibit sufficient power to reject dissimilar structures. In other words, we may conclude that both the atom pair and Euclidean distance methods are capable of choosing similar molecules from a collection of structurally diverse structures. This is in line with our earlier studies with various diverse sets of molecules [**10, 12, 15, 17, 18, 22, 23**].

The central paradigm of QSAR holds that similar structures usually have similar properties. To test this idea, we selected K-nearest neighbors ($K = 1 - 5$) for each molecule from a set of 113 mutagens and non-mutagens using the ED and AP methods and used the selected nearest neighbors in estimating mutagenicity. The results in Table 4 show that both methods lead to reasonably good estimates, although the AP method was superior to the ED method.

In conclusion, both the ED and AP methods, based on calculated graph theoretic structural invariants, did reasonably well in the selection of structural analogs and in the estimation of chemical properties based on nearest neighbors.

References

[1] H. Narumi, H. Hosoya, Topological Index and Thermodynamic Properties. II. Analysis of the Topological Factors on the Absolute Entropy of Acyclic Saturated Hydrocarbons, *Bull. Chem. Soc. Jpn.* **53** (1980), 1228–1237.

[2] M. Randić, Nonempirical Approaches to Structure-Activity Studies, *Int. J. Quantum Chem: Quant. Biol. Symp.***11** (1984), 137–153.

[3] R.E. Carhart, D.H. Smith, and R. Venkataraghavan, Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications, *J. Chem. Inf. Comput. Sci.* **25** (1985), 64–73.

[4] L.B. Kier, and L.H. Hall, Molecular Connectivity in Structure-Activity Analysis. Research Studies Press: Letchworth, Hertfordshire, U.K, 1986.

[5] D.H. Rouvray, and R.B. Pandey, The Fractal Nature, Graph Invariants and Physicochemical Properties of Normal Alkanes. *J. Chem. Phys.* **85** (1986), 2286–2290.

[6] S.C. Basak, Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **15** (1987), 605–609.

[7] S.C. Basak, G.J. Niemi, and G.D. Veith, In Computational Chemical Graph Theory, D.H. Rouvray, Ed.; NOVA: New York, 1990, pp. 235–277.

[8] N. Trinajstić, Chemical Graph Theory, Klein, D. J., and Randić, M., Eds.; CRC Press: Boca Raton, 1992.

[9] A.T. Balaban, S. Bertelsen, and S.C. Basak, New Centric Topological Indexes for Acyclic Molecules (Trees) and Substituents (Rooted Trees) and Coding of Rooted Trees. *Math. Chem.* **30** (1994), 55–72.

[10] S.C. Basak, S. Bertelsen, and G.D. Grunwald, Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.* **34** (1994), 270–276.

[11] S.C. Basak, G.D. Grunwald, and G.J. Niemi, Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure-Activity Relationships. In From Chemical Topology to Three Dimensional Molecular Geometry, Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73–116.

[12] M. Johnson, S.C. Basak, and G. Maggiora, A Characterization of Molecular Similarity Methods for Property Prediction. *Mathematical and Computer Modelling* **II** (1988), 630–635.

[13] S.C. Basak, V.R. Magnuson, G.J. Niemi, and R.R. Regal, Determining Structural Similarity of Chemicals using Graph-Theoretic Indices. *Discrete Appl. Math.* **19** (1988), 17–44.

[14] S.C. Basak, G.J. Niemi, and G.D. Veith, Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **7** (1991), 243–272.

[15] S.C. Basak, and G.D. Grunwald, Estimation of Lipophilicity from Structural Similarity. *New J. Chem.* **19** (1995), 231–237.

[16] S.C. Basak, and G.D. Grunwald, Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **31** (1995), 2529–2546.

[17] S.C. Basak, B.D. Gute, and G.D. Grunwald, Estimation of Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat. Chim. Acta* **69** (1996), 1159–1173.

[18] S.C. Basak, and B.D. Gute, Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal p-Hydroxylation of Anilines by Alcohol: A Molecular Similarity Approach. In Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health; Johnson, B. L., Xintaras, C., Andrews, J. S., Jr., Eds.; Princeton Scientific Publishing Co., Inc.: Princeton, NJ, 1997; pp 492–504.

[19] S.C. Basak, B.D. Gute, and G.D. Grunwald, Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals, Quantitative Structure-Activity Relationships. In Quantitative Structure-Activity Relationships in Environmental Sciences; Chen, F., Schuurman, G., Eds.; SETAC Press: Pensacola, FL, 1997; Vol. 7, Chapter 17, pp 245.

[20] S.C. Basak, B.D. Gute, and G.D. Grunwald, Use of Topostructural, Topochemical and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach, *J. Chem. Inf. Comput. Sci.* **37** (1997), 651–655.

[21] B.D. Gute, and S.C. Basak, Predicting Acute Toxicity (LC50) of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach, *SAR QSAR Environ. Res.* **7** (1997), 117–131.

[22] S.C. Basak, B.D. Gute, and G.D. Grunwald, Development and Applications of Molecular Similarity Methods using Nonempirical Parameters. *Mathl. Modelling Sci. Computing*, in press, 1999.

[23] S.C. Basak, and G.D. Grunwald, Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathl. Modelling Sci. Computing*, in press, 1999.

[24] B.D. Gute, G.D. Grunwald, and S.C. Basak, Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach, *SAR QSAR Environ. Res.* **10** (1999), 1–15.

[25] K. Balasubramanian, and S.C. Basak, Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **38** (1998), 367–373.

[26] S.C. Basak, B.D. Gute, and G.D. Grunwald, Prediction of Complement-Inhibitory Activity of Benzamidines Using Topological and Geometric Parameters. *J. Chem. Inf. Comput. Sci.* **39** (1999) 255–260.

[27] T. Yamaguchi, R. Hasegawa, A. Hagiwara, M. Hirose, K. Imaida, N. Ito, and T. Shirai, Results for 277 Chemicals in the Medium Term Liver Carcinogenesis Bioassay of Rats.

[28] H. Wiener, Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **69** (1947), 17–20.

[29] M. Randić, On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **97** (1975), 6609–6615.

[30] D. Bonchev, and N. Trinajstić, Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **67** (1977), 4517–4533.

[31] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak, Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *J. Comput. Chem.* **5** (1984), 581–588.

[32] S.C. Basak, A.B. Roy, and J.J. Ghosh, Study of the Structure-Function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In Proceedings of the 2nd International Conference on Mathematical Modelling, Avula, X. J. R., Bellman, R., Luke, Y. L., and Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, Missouri, 1980; Vol. II, pp. 851–856.

[33] S.C. Basak, and V.R. Magnuson, Molecular Topology and Narcosis: A Quantitative Structure-Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim. Forsch.* **33** (1983), 501–503.

[34] A.B. Roy, S.C. Basak, D.K. Harriss, and V.R. Magnuson, Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In Mathematical Modelling in Science and Technology, Avula, X. J. R., Kalman, R. E., Lipais, A. I., and Rodin, E. Y., Eds.; Pergamon Press: New York, 1984, pp. 745–750.

[35] A.T. Balaban, Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **89** (1982), 399–404.

[36] A.T. Balaban, Topological Indices Based on Topological Distances in Molecular Graphs. *Pure & Appl. Chem.* **55** (1983), 199–206.

[37] A.T. Balaban, Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking into account Periodicities of Element Properties. *Math. Chem. (MATCH)* **21** (1986), 115–122.

[38] S.C. Basak, D.K. Harriss, and V.R. Magnuson, POLLY 2.3: Copyright of the University of Minnesota, 1988.

[39] C.E. Shannon, A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27** (1948), 379–423.

[40] N. Rashevsky, Life, Information Theory and Topology. *Bull. Math. Biophys.* **17** (1955), 229–235.

[41] R. Sarkar, A.B. Roy, and R.K. Sarkar, Topological Information Content of Genetic Molecules — I. *Math. Biosci.* **39** (1978), 299–312.

[42] V.R. Magnuson, D.K. Harriss, and S.C. Basak, Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications. In Studies in Physical and Theoretical Chemistry, King, R. B., Ed.; Elsevier: Amsterdam, 1983, pp. 178–191.

[43] S.C. Basak, and G.D. Grunwald, APProbe: Copyright of the University of Minnesota, 1993.

[44] SAS Institute Inc, in: SAS/STAT User's Guide, Release 6.03 Edition (SAS Institute Inc., Cary, NC, 1988 p. 751.

[45] C.L. Wilkins, and M. Randić, A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations. *Theor. Chim. Acta* **58** (1980), 45–68.

Natural Resources Research Institute, 5013 Miller Trunk Hwy., Duluth, MN, 55811 USA

E-mail address: sbasak@wyle.nrri.umn.edu

Natural Resources Research Institute, 5013 Miller Trunk Hwy., Duluth, MN, 55811 USA

E-mail address: bgute@wyle.nrri.umn.edu

APPENDIX 1.14 Use of statistical and neural net approaches in predicting toxicity of chemicals

Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals

Subhash C. Basak,*,[†] Gregory D. Grunwald,[†] Brian D. Gute,[†] Krishnan Balasubramanian,[‡] and David Opitz[§]

Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, Department of Chemistry and Biochemistry, Physical Sciences Building, Arizona State University, D-106 Tempe, Arizona 85287-1604, and Department of Computer Science, Social Science Building, University of Montana, Missoula, Montana 5981

Received September 9, 1999

Hierarchical quantitative structure–activity relationships (H-QSAR) have been developed as a new approach in constructing models for estimating physicochemical, biomedicinal, and toxicological properties of interest. This approach uses increasingly more complex molecular descriptors in a graduated approach to model building. In this study, statistical and neural network methods have been applied to the development of H-QSAR models for estimating the acute aquatic toxicity (LC_{50}) of 69 benzene derivatives to *Pimephales promelas* (fathead minnow). Topostructural, topochemical, geometrical, and quantum chemical indices were used as the four levels of the hierarchical method. It is clear from both the statistical and neural network models that topostructural indices alone cannot adequately model this set of congeneric chemicals. Not surprisingly, topochemical indices greatly increase the predictive power of both statistical and neural network models. Quantum chemical indices also add significantly to the modeling of this set of acute aquatic toxicity data.

1. INTRODUCTION

An important aspect of modern toxicology research is the prediction of toxicity of xenobiotics and environmental pollutants from their molecular structure.[1–13] The potential toxicity of a chemical is normally assessed on the basis of a wide variety of relevant physical and biological properties. Table 1 provides a partial list of such properties. Risk assessors use these kinds of toxicological indicators to estimate the potential risk posed by a given compound, using simpler properties relevant to a chemical's toxicity to make more complex assessments relevant to human and environmental health. However, the Toxic Substances Control Act (TSCA) Inventory currently includes about 80 000 chemicals, most of which do not have data for the toxicologically relevant properties mentioned in Table 1. In fact, roughly 50% of these chemicals do not have any experimental property data at all.[14] Worldwide, more than 16.7 million distinct organic and inorganic chemicals are known, as is evident from the number of entries in the Chemical Abstract Service (CAS) inventory.[15] For many of these chemicals we do not have the data necessary for risk assessment. Additionally, modern combinatorial chemistry techniques have led to the production of vast libraries of chemicals at a very rapid rate. Most of these substances have none of the test data needed for their hazard estimation.

Recently there have been efforts by the chemical industry and government agencies to develop reliable databases of properties that will be used for hazard estimation.[16] This effort, although commendable, falls short of the need; and the picture will remain so in the foreseeable future. In the area of molecular biology, innovative techniques are emerging where specially engineered cell lines can be used to detect the activity or toxicity of chemicals to the genetic system.[17–19] Effects of chemicals on the pattern of cellular proteins, analyzed by proteomics technology, are being used to detect their potential toxic effects.[20–22] Such methods are faster than the traditional in vivo test methods, and it is possible that they could be developed to the point where they will replace or significantly decrease the need for whole-animal screening methods. At present, neither the available test data nor the combination of in vitro toxicity testing methods provides adequate resources for hazard assessment.

Table 1. Physicochemical and Biological Properties Relevant to the Assessment of toxicity

physicochemical	biological
molar volume	receptor binding (K_D)
boiling point	Michaelis constant (K_m)
melting point	inhibitor constant (K_i)
vapor pressure	biodegradation
aqueous solubility	bioconcentration
dissociation constant (pK_a)	alkylation profile
partition coefficient	metabolic profile
octanol–water (log *P*)	chronic toxicity
air–water	carcinogenicity
sediment–water	mutagenicity
reactivity (electrophile)	acute toxicity
	LD_{50}
	LC_{50}

Quantitative structure–activity/–toxicity relationship (QSAR/QSTR) models have emerged as useful tools to handle the data gap in toxicology and pharmacology.[1–13,22–26] QSAR models can be used to estimate complex properties of chemicals from simpler experimental or computed proper-

* To whom all correspondence should be addressed. Telephone: (218) 720-4230. Fax: (218) 720-4328. E-mail: sbasak@nrri.umn.edu.

[†] University of Minnesota, Duluth.

[‡] Arizona State University.

[§] University of Montana.

10.1021/ci9901136 CCC: $19.00
Published on Web 06/22/2000

ties. In view of the fact that most chemicals in commerce and environmental pollutants have very little test data, it would be desirable if we could develop toxicologically relevant QSARs from properties that can be calculated directly from a chemical's structure. In some of our recent papers we have developed a novel hierarchical QSAR (H-QSAR) approach where four classes of theoretical molecular descriptors, viz., topostructural, topochemical, geometrical, and quantum chemical parameters, have been used sequentially in the formulation of H-QSAR models for predicting physical, biomedicinal, and toxicological properties.[1,3,6,8,23−26]

Most of our H-QSARs are based on linear statistical methods such as multiple linear regression, principal components analysis (PCA), and variable clustering. Such methods yield useful models, but they suffer from the limitation that in some cases the relationship between a molecular descriptor and toxicity may be intrinsically nonlinear. In such cases, the use of linear statistical methods may not result in the best models. Therefore, in this paper, we have carried out a comparative study of multiple regression vis-à-vis neural net methods in predicting the acute aquatic toxicity (LC_{50}) of a set of 69 benzene derivatives.

2. METHODS

2.1. Toxicity Database. The utility of this approach of generating numerous hierarchical theoretical descriptors of compounds was tested on a set of acute aquatic toxicity (LC_{50}) data for 69 benzene derivatives. The data were taken from a study by Hall et al.,[12] who collected acute aquatic toxicity data measured in fathead minnow (*Pimephales promelas*). These data were compiled from eight other literature sources and included some original work which was conducted at the U. S. Environmental Protection Agency Environmental Research Laboratory (USEPA-ERL) in Duluth, MN. This set of chemicals was composed of benzene and 68 substituted benzene derivatives. According to the authors, these benzene derivatives were tested using methodologies comparable to their own 96-h fathead minnow toxicity test system. The derivatives chosen for this study (see Table 2) have seven different substituent groups that are present in at least six of the molecules: chloro-, bromo-, nitro-, methyl-, methoxyl-, hydroxyl-, and amino-.

2.2. Calculation of Topological Indices. The complete set of topological indices (TIs) used in this study, both topostructural and topochemical, have been calculated using POLLY 2.3 and other software developed by Basak et al.[27] These indices include the Wiener index,[28] the connectivity indices developed by Randić,[29] higher order connectivity indices formulated by Kier and Hall,[30] bonding connectivity indices defined by Basak et al.,[31] a set of information theoretic indices defined on the distance matrices of simple molecular graphs,[32,33] a set of parameters derived on the neighborhood complexity of hydrogen-filled molecular graphs,[34−36] and Balaban's J indices.[37−39] Table 3 provides the symbols of the topological indices and brief definitions.

The set of TIs was divided into two distinct subsets: topostructural indices (TSI) and topochemical indices (TCI). TSIs are topological indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors such as hybridiza-

Table 2. Experimental and Estimated Acute Aquatic Toxicity Data for 69 Benzene Derivatives, Expressed as $-\log(LC_{50})$ for the Linear Regression Model (LR) and the Neural Network Model Using the 23 Parameters Selected by Variable Clustering

compound	expt	LR	NN
benzene	3.40	3.42	3.65
bromobenzene	3.89	3.77	3.79
chlorobenzene	3.77	3.75	3.77
phenol	3.51	3.38	3.51
toluene	3.32	3.66	3.62
1,2-dichlorobenzene	4.40	4.29	4.30
1,3-dichlorobenzene	4.30	4.37	4.12
1,4-dichlorobenzene	4.62	4.51	4.27
2-chlorophenol	4.02	3.79	3.91
3-chlorotoluene	3.84	3.88	3.79
4-chlorotoluene	4.33	3.87	3.76
1,3-dihydroxybenzene	3.04	3.43	3.53
3-hydroxyanisole	3.21	3.33	3.45
2-methylphenol	3.77	3.64	3.67
3-methylphenol	3.29	3.60	3.58
4-methylphenol	3.58	3.53	3.55
4-nitrophenol	3.36	3.61	3.76
1,4-dimethoxybenzene	3.07	3.28	3.51
1,2-dimethylbenzene	3.48	3.93	3.91
1,4-dimethylbenzene	4.21	3.87	3.68
2-nitrotoluene	3.57	3.66	3.81
3-nitrotoluene	3.63	3.53	3.71
4-nitrotoluene	3.76	3.49	3.68
1,2-dinitrobenzene	5.45	5.24	4.99
1,3-dinitrobenzene	4.38	4.18	4.19
1,4-dinitrobenzene	5.22	4.94	4.85
2-methyl-3-nitroaniline	3.48	3.79	3.88
2-methyl-4-nitroaniline	3.24	3.51	3.75
2-methyl-5-nitroaniline	3.35	3.68	3.86
2-methyl-6-nitroaniline	3.80	3.84	3.79
3-methyl-6-nitroaniline	3.80	3.78	3.62
4-methyl-2-nitroaniline	3.79	3.80	3.66
4-hydroxy-3-nitroaniline	3.65	3.61	3.58
4-methyl-3-nitroaniline	3.77	3.73	3.72
1,2,3-trichlorobenzene	4.89	4.89	5.04
1,2,4-trichlorobenzene	5.00	5.04	4.83
1,3,5-trichlorobenzene	4.74	5.11	4.78
2,4-dichlorophenol	4.30	4.33	4.47
3,4-dichlorotoluene	4.74	4.26	4.28
2,4-dichlorotoluene	4.54	4.36	4.44
4-chloro-3-methylphenol	4.27	3.87	4.07
2,4-dimethylphenol	3.86	3.76	3.72
2,6-dimethylphenol	3.75	3.80	3.84
3,4-dimethylphenol	3.90	3.80	3.79
2,4-dinitrophenol	4.04	4.14	4.01
1,2,4-trimethylbenzene	4.21	4.09	3.87
2,3-dinitrotoluene	5.01	5.20	5.28
2,4-dinitrotoluene	3.75	4.10	4.33
2,5-dinitrotoluene	5.15	4.84	4.72
2,6-dinitrotoluene	3.99	4.41	4.63
3,4-dinitrotoluene	5.08	5.11	5.09
3,5-dinitrotoluene	3.91	4.05	4.16
1,3,5-trinitrobenzene	5.29	5.37	5.32
2-methyl-3,5-dinitroaniline	4.12	4.13	4.23
2-methyl-3,6-dinitroaniline	5.34	4.80	4.54
3-methyl-2,4-dinitroaniline	4.26	4.28	4.20
5-methyl-2,4-dinitroaniline	4.92	4.14	4.02
4-methyl-2,6-dinitroaniline	4.21	4.67	4.58
5-methyl-2,6-dinitroaniline	4.18	4.80	4.78
4-methyl-3,5-dinitroaniline	4.46	4.34	4.43
2,4,6-tribromophenol	4.70	4.89	5.47
1,2,3,4-tetrachlorobenzene	5.43	5.62	5.56
1,2,4,5-tetrachlorobenzene	5.85	5.80	5.61
2,4,6-trichlorophenol	4.33	4.79	4.96
2-methyl-4,6-dinitrophenol	5.00	4.21	4.16
2,3,6-trinitrotoluene	6.37	6.36	5.81
2,4,6-trinitrotoluene	4.88	5.16	5.42
2,3,4,5-tetrachlorophenol	5.72	5.36	5.58
2,3,4,5,6-pentachlorophenol	6.06	6.03	5.83

Table 3. Symbols, Definitions, and Classifications of Topological, Geometrical, and Quantum Chemical Parameters

	Topostructural
$\underline{I}_D^W$	information index for the magnitudes of distances between all possible pairs of vertexes of a graph
$\bar{I}_D^W$	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	degree complexity
H^V	graph vertex complexity
$\underline{H}^D$	graph distance complexity
$\overline{\mathrm{IC}}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	a Zagreb group parameter = sum of square of degree over all vertexes
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertexes
${}^h\chi$	path connectivity index of order $h = 0-6$
${}^h\chi_C$	cluster connectivity index of order $h = 3, 5$
${}^h\chi_{Ch}$	chain connectivity index of order $h = 6$
${}^h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
P_h	no. of paths of length $h = 0-10$
J	Balaban's J index based on distance
	Topochemical
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertexes
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
SIC_r	structural information content for r^{th} ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) order neighborhood of vertexes in a hydrogen-filled graph
${}^h\chi^b$	bond path connectivity index of order $h = 0-6$
${}^h\chi^b{}_C$	bond cluster connectivity index of order $h = 3, 5$
${}^h\chi^b{}_{Ch}$	bond chain connectivity index of order $h = 6$
${}^h\chi^b{}_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	valence path connectivity index of order $h = 0-6$
${}^h\chi^v{}_C$	valence cluster connectivity index of order $h = 3, 5$
${}^h\chi^v{}_{Ch}$	valence chain connectivity index of order $h = 6$
${}^h\chi^v{}_{PC}$	valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
	Geometrical
V_W	van der Waals volume
${}^{3D}W$	3D Wiener no. for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3D Wiener no. for the hydrogen-filled geometric distance matrix
	Quantum Chemical
E_{HOMO}	energy of the highest occupied molecular orbital
E_{HOMO1}	energy of the second highest occupied molecular orbital
E_{LUMO}	energy of the lowest unoccupied molecular orbital
E_{LUMO1}	energy of the second lowest unoccupied molecular orbital
ΔH_f	heat of formation
μ	dipole moment

tion states of atoms and number of core/valence electrons in individual atoms. TCIs are parameters that quantify information regarding the topology (connectivity of atoms), as well as specific chemical properties of the atoms and bonds comprising a molecule. TCIs are derived from weighted molecular graphs where each vertex (atom) is properly weighted with relevant chemical/physical properties. Table 3 shows the division of the topological indices into topostructural and topochemical indices.

2.3. Calculation of Geometrical Indices. The geometrical indices include the three-dimensional (3D) Wiener numbers for hydrogen-filled and hydrogen-suppressed molecular structures and van der Waals volume. van der Waals volume, V_W, was calculated using SYBYL 6.4 from Tripos Associates, Inc.[40] The 3D Wiener numbers were calculated by SYBYL using an SPL (Sybyl Programming Language) program developed in our laboratory. Calculation of the 3D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3D coordinates for the atoms were determined using CONCORD 3.2.1.[41] The symbols and definitions of the geometrical indices are included in Table 3.

2.4. Quantum Chemical Parameters. Quantum chemical parameters were calculated using the Austin Model version one (AM1) semiempirical Hamiltonian. These parameters were calculated using MOPAC 6.00 in the SYBYL interface.[42] Brief definitions and symbols for the quantum chemical parameters used in this study are included in Table 3.

2.5. Statistical Analysis and Hierarchical QSAR. Initially, all topological indices were transformed by the natural logarithm of the index plus one. This was done to scale the indices, since some may be several orders of magnitude greater than others, while other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency; the addition of one was unnecessary.

The set of 86 topological indices was then partitioned into the two distinct sets: topostructural indices (35) and topochemical indices (51). The sets of topostructural and topochemical indices were then divided into subsets, or clusters, based on the correlation matrix using the SAS variable clustering procedure (VARCLUS)[43] to further reduce the number of independent variables for use in model construction. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional.

From each cluster, the index most correlated with the cluster was selected for modeling, as well as any indices that were poorly correlated with their cluster ($R^2 < 0.70$). These indices were then used in the modeling of the acute aquatic toxicity of benzene derivatives in fathead minnow. The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure resulted in a set of five topostructural indices and a set of nine topochemical indices.

Reducing the number of independent variables is critical when attempting to model small data sets using linear statistical methods. The smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors). A study by Topliss and Edwards[44] has shown that for a set with about 70 dependent variables (observations), no more than 40 independent variables may be used while keeping the probability of chance correlations below 1%. This number is dependent on the actual correlation achieved in the modeling process; higher correlation results in a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cutoff of 40 independent variables. In fact, the total number of descriptors which will be used for model construction and estimation is 23, well within the bounds of the Topliss and Edwards criteria.[44]

Regression modeling was accomplished using the SAS procedure REG[43] on four distinct sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest parameters, the TSIs. After using the TSIs to model the activity, the next level of parameters are added. To the indices included in the best TSI model, we add all of the TCIs and proceed to model the activity using these parameters. Likewise, the indices included in the best model from this procedure are combined with the indices from the next complexity level, the geometrical indices, and modeling is conducted once again. Finally, the best model utilizing TSIs, TCIs, and geometrical indices is combined with the quantum chemical parameters to develop the final model in the hierarchy.

Additionally, the entire set of 95 descriptors (topostructural, topochemical, geometrical, and quantum chemical) was subjected to the variable clustering procedure and a reduced set of independent variables was used in constructing a QSAR model. This varies from the other approach in that the indices were clustered as one set, rather than as four distinct sets, and resulted in a somewhat different set of variables. This was done to determine if there is any advantage in final model predictive power between model development based on the H-QSAR approach versus the "kitchen sink" approach, i.e., using the entire descriptor set in order to find the "best" model.

2.6. Neural Network Methods. Using neural networks, we studied two classes of approaches for modeling toxicity: (1) giving all the descriptors to a learning algorithm (neural network in this case) and (2) reducing the feature set before giving the (reduced) feature set to a learning algorithm. Results for our approaches are from leave-one-out experiments (i.e., 69 training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (68 in this case) to the learning algorithms for training. This process is repeated 69 times so that each example is a part of the test set once and only once. Leave-one-out tests *generalization* accuracy of a learner, whereas training set accuracy tests only the learner's ability to memorize. Generalization error from the test set is the true test of accuracy and is what we report here.

First we trained neural networks using all 95 parameters: 35 TSI, 51 TCI, 3 geometrical, and 6 quantum chemical parameters. The networks contained 15 hidden units and were trained for 1000 epochs. Each input parameter was normalized to a value between 0 and 1 before training. Additional parameter settings for the neural networks included a learning rate of 0.05, a momentum term of 0.1, and weights initialized randomly between −0.25 and +0.25.

For our next experiment, we used a smaller set of 23 independent variables divided further into the four levels of the hierarchy. The 23 independent variables included the 5 topostructural and 9 topochemical parameters provided by the variable clustering technique (see section 3.1 for a list of the indices) combined with the 3 geometrical and 6 quantum chemical parameters described in Table 3. The parameter settings for these networks were the same as the settings for the other neural network experiment mentioned above.

3. RESULTS

3.1. Results of Statistical Regression Procedures. The variable clustering of the topostructural indices resulted in the retention of five indices: M_1, $\overline{\mathrm{IC}}$, O, P_8, P_9. All-subsets regression resulted in the selection of a four-parameter model to estimate $-\log(\mathrm{LC}_{50})$ with an explained variance (R^2) of 45.3% and a standard error (s) of 0.58. While this is an unsatisfactory model, the indices were retained and combined with the topochemical indices in the second step of model development. The second step combined the 4 indices used in the first tier model with the 9 topochemical indices selected in the variable clustering procedure: SIC_0, SIC_1, SIC_4, CIC_0, ${}^{2}\chi^{\mathrm{b}}$, ${}^{5}\chi^{\mathrm{b}}{}_{\mathrm{C}}$, ${}^{5}\chi^{\mathrm{v}}{}_{\mathrm{C}}$, ${}^{6}\chi^{\mathrm{v}}{}_{\mathrm{PC}}$, J^{X}. Again, all-subsets regression was conducted resulting in a four-parameter model with an explained variance (R^2) of 78.3% and a standard error (s) of 0.36. The 4 indices from the second tier model were combined with the three geometric parameters: ${}^{3\mathrm{D}}W_{\mathrm{H}}$, ${}^{3\mathrm{D}}W$, V_{W}. This resulted in a four-parameter model that replaced the topochemical index CIC_0 with the geometric parameter ${}^{3\mathrm{D}}W_{\mathrm{H}}$. This model had an explained variance (R^2) of 79.2% and a standard error (s) of 0.36. The final step in the hierarchical method combined the four parameters from the third tier model with the semiempirical quantum chemical parameters: E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔH_f, μ. This set of 10 indices led to a seven-parameter model with an explained variance (R^2) of 86.3% and a standard error (s) of 0.30. This model retained all indices from the third model and added three of the AM1 quantum chemical parameters. Our final model, using indices selected from a variable clustering of the entire set of 95 indices resulted in a seven-parameter model including three topostructural indices (${}^{0}\chi$, P_9, $\overline{\mathrm{IC}}$), one topochemical index (${}^{5}\chi^{\mathrm{v}}$), one geometrical index (${}^{3\mathrm{D}}W_{\mathrm{H}}$), and two quantum chemical descriptors (ΔH_f, μ). This model had an explained variance (R^2) of 86.1% and a standard error (s) of 0.30.

Table 4. Relative Effectiveness of Statistical and Neural Network Methods in Estimating the Acute Aquatic Toxicity of 69 Benzene Derivatives

	neural networks		linear regression	
model	R_c^2	s	R_c^2	s
TSI	0.299	0.63	0.366	0.629
+ TCI	0.619	0.47	0.754	0.392
+ 3D	0.656	0.44	0.763	0.384
+ QC	0.770	0.36	0.825	0.339
all 95 indices	0.758	0.37	0.827	0.337

Leave-one-out analysis was conducted on all models for purposes of comparison with the results from the neural networks. The resulting values for cross-validated R^2 (R_c^2) and standard error (s) are reported in Table 4.

3.2. Results of the Neural Network Procedures. The first approach incorporating all 95 parameters, obtained a test-set correlation coefficient between predicted toxicity and measured toxicity (explained variance) of $R^2 = 0.868$ and a standard error of 0.29. The second approach utilizes the hierarchical method of grouping descriptors resulted in four models, one for each level of the hierarchy. The results from the leave-one-out analysis of these four models, as well as those for the linear statistical models are summarized in Table 4. Table 2 presents the experimental acute aquatic toxicity ($-\log[LC_{50}]$) values for the 69 benzene derivatives as well as the values estimated by the best statistical model and the best neural network model, both of which resulted from the fourth H-QSAR model.

4. DISCUSSION

The results show that both statistical and neural network models give acceptable estimates for the toxicity of the 69 benzene derivatives studied in this paper. As can be clearly seen from the comparative results in Table 4, there are two points in the hierarchical approach in which there are significant improvements in modeling the data. The addition of the topochemical indices increases the variance explained in both the statistical and neural network models by 30−40% with a consequent drop in the standard error of the calculations as well. Addition of the quantum chemical parameters also creates a significant increase in the efficacy of both models, a 6.2% increase in the variance explained for the statistical model and an 11.4% increase for the neural network model.

It is interesting to note that the neural network model using the subset of 23 inputs selected in part by the VARCLUS procedure gave slightly better results as compared to the network developed using all 95 input variables. This could be the result of filtering out redundant, or nearly redundant, parameters from the set of independent variables.

Further work on the relative utility of statistical vis-à-vis neural network methods is necessary to determine which types of models are best suited to the estimation of chemical toxicity.

ACKNOWLEDGMENT

This paper is contribution no. 270 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, Grant IRI-9734419 from the National Science Foundation, and a MONTS grant from the University of Montana.

REFERENCES AND NOTES

(1) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Reading, U.K., 1999; pp 675−696.

(2) Basak, S. C. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990; pp 83−103.

(3) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *QSAR in Environmental Sciences*, Vol. 7; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1998; pp 245−261.

(4) Basak, S. C.; Gute, B. D. In *Discrete Mathematical Chemistry*; Hansen, P., Paradis, N., Eds.; DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 51; American Mathematical Society: Providence, RI, 2000; pp 9−24.

(5) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of the mutagenicity of chemicals from theoretical structural parameters: A hierarchical approach. *SAR QSAR Environ. Res.* **1999**, *10*, 117−129.

(6) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.

(7) Basak, S. C.; Gute, B. D. Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1−21.

(8) Gute, B. D.; Basak, S. C. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.

(9) Mushrush, G. W.; Basak, S. C.; Slone, J. E.; Beal, E. J.; Basu, S.; Stalick, W. M.; Hardy, D. R. Computational study of the environmental fate of selected aircraft deicing compounds. *J. Environ. Sci. Health* **1997**, *A32* (8), 2201−2211.

(10) Basak, S. C.; Grunwald, G. D. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. *Chemosphere* **1995**, *31*, 2529−2546.

(11) Basak, S. C.; Grunwald, G. D. In *Proceeding of the XVI International Cancer Congress*; Rao, R. S., Deo, M. G., Sanghui, L. D., Eds.; Monduzzi: Bologna, Italy, 1995; p 413.

(12) Hall, L.; Kier, L.; Phipps, G. Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.* **1984**, *3*, 355−365.

(13) Gombar, V. K.; Enslein, K.; Blake, B. W. Assessment of developmental toxicity potential of chemicals by quantitative structure-toxicity relationship models. *Chemosphere* **1995**, *31*, 2499−2510.

(14) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health. Perspect.* **1990**, *87*, 183−197.

(15) CAS. The latest CAS registry number and substance count. http://www.cas.org/cgi-bin/regreport.pl, 2000.

(16) Johnson, J. Pact triggers tests: Thousands of chemicals may be tested under toxicity screening program. *Chem. Eng. News* **1998**, *76*, 19−20.

(17) Chen, J. J.; Wu, R.; Yang, P. C.; Huang, J. Y.; Sher, Y. P.; Han, M. H.; Kao, W. C.; Lee, P. J.; Chiu, T. F.; Chang, F.; Chu, Y. W.; Wu, C. W.; Peck, K. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* **1998**, *51*, 313−324.

(18) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467−470.

(19) De Risi, J.; Penland, L.; Brown, P. O.; Bittner, M. L.; Meltzer, P. S.; Ray, M.; Chen, Y.; Su, Y. A.; Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **1996**, *14*, 457−460.

(20) Witzmann, F. A.; Fultz, C. D.; Grant, R. A.; Wright, L. S.; Kornguth, S. E.; Siegel, F. L. Differential expression of cytosolic proteins in the rat kidney cortex and medulla: Preliminary proteomics. *Electrophoresis* **1998**, *19*, 2491−2497.

(21) Anderson, N. L.; Esquer-Blasco, R.; Richardson, F.; Foxworthy, P.; Eacho, P. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharm.* **1996**, *137*, 75−89.

(22) Lake, B. G.; Lewis, D. F. V.; Gray, T. J. B.; Beamand, J. A. Structure-activity relationships for induction of peroxysomal enzyme activities in primary rat hepatocyte cultures. *Toxicol. in Vitro* **1993**, *7*, 605−614.

(23) Basak, S. C.; Gute, B. D.; Grunwald, G. D.; Opitz, D. W.; Balasubramanian, K. In *Predictive Toxicology of Chemicals: Experiences and*

Impact of AI Tools-Papers from the 1999 AAAI Symposium; AAAI Press: Menlo Park, CA, 1999; pp 108−111.

(24) Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 255−260.

(25) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.

(26) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060.

(27) Basak, S.; Harriss, D.; Magnuson, V. *POLLY 2.3*; University of Minnesota: Duluth, MN, 1988.

(28) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(29) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(30) Kier, L.; Hall, L: *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Hertfordshire, U.K., 1986.

(31) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.

(32) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.* **1984**, *5*, 581−588.

(33) Bonchev, D.; Trinajstić, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(34) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*, Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri − Rolla: rolla, MO, 1980; p 851.

(35) Roy, A. B.; Basak, S. C.: Harriss, D. K.; Magnuson, V. R. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; p 745.

(36) Basak, S. C.; Magnuson, V. R. Molecular topology and narcosis. *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501−503.

(37) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(38) Balaban, A. T. Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.

(39) Balaban, A. T. Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.

(40) *SYBYL Version 6.4.*; Tripos Associates, Inc.: St. Louis, MO, 1998.

(41) *CONCORD Version 3.2.1.*; Tripos Associates, Inc.: St. Louis, MO, 1998.

(42) Stewart, J. J. P. *MOPAC 6.00*, QCPE #455; Frank J. Seiler Research Laboratory, U.S. Air Force Academy: Colorado Springs, CO, 1990.

(43) *SAS/STAT User's Guide*, 6.03 ed.; SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp 773−875, 949−965.

CI9901136